Relationship between Hypertension and Psychopathologies

Anton Barrera Mora

March 2023

Contents

ENGLISH

Data Mining Pilot Project Proposal: Relationship between Hypertension and Psychopathologies

Phase 1: Understanding the Business

Overall objective of the phase:

Definition of the research problem and establishment of project objectives.

Problem:

The mental health department of a geriatric company has identified indicators that suggest a link between hypertension and worsening mental disorders, as well as the occurrence of psychotic episodes and crises.

Objectives:

Statistical refutation: Analyze historical data from the company to establish statistical significance between hypertension and behavioral problems, and determine correlation, causation, or dismiss the theory.

Prevention and improvement of care: Implement measures through a model that allows early detection, based on thresholds of physiological measurements and medical histories, enabling the establishment of preventive measures.

Product:

Foundation document for the creation of a machine learning project. This would enable improvements in the healthcare management software by detecting risk parameters and triggering early alerts. The document would provide a breakdown of relevant variables related to the problem and specify general and specific blood pressure thresholds that may affect each patient. It would be a **predictive model.**

Tasks:

1. Definition of the study population - clients residing in any nursing home or facility owned by the company.
2. Collection of primary data from various sources within the company.
3. Collection of secondary data related to the problem, whether collected by other institutions or from research studies - in this project, it is proposed to avoid tertiary sources or data from third parties that have not been directly collected by the publisher.
4. Identification of relevant variables for analysis.

Success criteria:

Establishment of correlation or causation between mental health and hypertension, considering that hypertension can be both a factor influencing the development of a mental illness and a factor exacerbating symptoms once a psychopathology is developed. In essence, we will examine the role played by hypertension and its relationship with mental or neurological disorders.

Design of a baseline model for the development of a module or machine learning program, or a similar solution, deployed within the company.

Phase 2: Understanding the Data

Overall objective of the phase:

Identify and understand the data necessary for the development of the data mining project.

Objectives:

Discard datasets unrelated to the project - e.g., purchase records, family visits, hygiene, etc.

Understand the characteristics and structure of the data necessary for the project - file formats and/or types, location, accessibility, security, number of records, attributes, etc.

Product:

Comprehensive descriptive document of the available data - quality, source, age, relevance, roles in the project, etc.

Tasks:

Exploratory data analysis to draw initial conclusions - number of records, number of variables, typology, distribution of values, errors (incorrect format, missing values, empty fields, whitespace), values (outliers, inappropriate formats).

Phase 3: Data Preparation

Overall objective of the phase:

Prepare the data based on the aspects identified in Phase 2 and make it suitable for analysis.

Products:

Cleaned and transformed dataset for further analysis.

Changelog or data file change log.

Tasks:

Dimensionality reduction.

Reduction and filtering of attributes and features - e.g., exclude dates that fall outside the study period or subjects below the specified age in the project.

Removal of outliers that disrupt the model's accuracy - e.g., extremely high or low blood pressure values that are unlikely or even impossible to measure and need to be removed from the dataset to maintain model accuracy.

Dimensionality reduction using techniques such as Principal Component Analysis (PCA) or Canonical Correlation Analysis (CCA) to reduce the number of variables to a few factors that explain most of the data variability - e.g., reducing daily records to representative blood pressure values.

Identification of highly correlated variables using correlation matrix and iterative feature elimination (RFE) techniques.

Other outlier detection techniques such as multivariate analysis, box plots, and whisker plots.

Sampling and condensing to reduce data size.

Creation of new variables by combining or deriving existing variables that may be more useful for analysis - e.g., using food intake and fluid intake records to create a new variable 'Hydration'.

Data transformation:

Data normalization.

Data scaling - e.g., changing meters to kilometers in the physical activity records.

Normalization by maximum value - e.g., normalizing residents' age by dividing each resident's age by the maximum age, resulting in a scale from 0 to 1.

Normalization by difference - e.g., normalizing daily antipsychotic doses (Risperidone) by calculating the normalized dose using the formula: normalized dose = (dose - minimum dose) / (maximum dose - minimum dose), resulting in a value between 0 and 1.

Normalization by standard deviation - e.g., normalizing the "Risperidone dose" variable by dividing it by the standard deviation of the variable.

Encoding categorical variables by assigning numerical values to enable statistical analysis - e.g., transforming physical activity categories (sedentary, moderately active, very active) into numerical values such as 1, 2, and 3, respectively.

Data cleaning:

Identification of biases - ideological, gender biases, etc. - in the data and/or samples - observer bias, interpretation bias, confirmation bias.

Example of ideological bias: A researcher with a strong political stance wants to demonstrate that hypertension is strongly related to lack of access to medical care, selectively choosing or interpreting data to support their hypothesis while ignoring other important variables.

Example of gender bias: A researcher may be more interested in the relationship between the disease and men, focusing mainly on this group and ignoring important factors in hypertension among women, such as menopause, pregnancy, or hormonal contraception.

Example of observer bias: A researcher is interested in exploring the relationship between stress and hypertension due to cases within their family. Consequently, they primarily observe cases where stress is present with hypertension.

Example of interpretation bias: A researcher may find a correlation between hypertension and age but mistakenly interpret that hypertension is caused by age, rather than age being a risk factor for hypertension.

Example of confirmation bias: A researcher may have a preconceived hypothesis about the relationship between the disease and salt consumption, actively seeking data or analyses that support their hypothesis while disregarding other contributing factors such as genetics or overall lifestyle.

Ensuring the reliability, originality, comprehensibility, currency, and validity of data sources to be used.

Proper data cleaning considering aspects such as:

Incomplete data - e.g., missing blood pressure measurements on certain days. Regression imputation is not appropriate for this project, so we decide to remove the incomplete days.

Incorrect data - e.g., a date of 1705 in the "date" field related to a birthdate, which must be an error.

Data integrity or lack of precision, completeness, consistency, and/or credibility - e.g., different identifier formats as primary keys, blood pressure measurements with two digits, etc.

Duplicate data - e.g., the same instance with the same username repeated in three records, so we decide to merge or integrate them.

Removal of white spaces - e.g., removing double spaces in compound names.

Orthographic correction and correction of spelling errors.

Ensuring compatibility of datasets to work with foreign keys and enable data merging.

Fixing incorrect data formats - e.g., using text format instead of the appropriate date format for working with dates.

Phase 4: Modeling

Overall objective of the phase:

Construct a data mining model and evaluate its performance.

Product:

Data mining model on hypertension and psychopathologies in the elderly population, with the ability to predict behavioral problems based on various parameters and anticipate hypertension issues based on specific psychophysical test values.

Model evaluation.

Tasks:

Selection of modeling techniques. This involves selecting algorithms that allow for the prediction of hypertension probability based on variables such as age, gender, BMI, anxiety, and the risk of experiencing psychotic episodes based on medication dosage, weather conditions, gender, age, and specific blood pressure thresholds. Decision tree algorithms are used to classify new observations accurately by constructing a tree where each node represents a decision variable and each branch represents a possible response to that variable. The goal is to find the decision variable that best separates the observations based on their class - the variable that provides the most accurate classification. Different decision variables are evaluated, and the one that yields the highest benefit in terms of accuracy is selected.

Practical example in R:

Suppose we want to predict whether a person has hypertension based on their age, gender, body mass index (BMI), and anxiety level:

```
# Load the library
library(rpart)

# Load the ANTONSET dataset derived from previous phases
ANTONSET <- read.csv

# Create a column for the target variable (hypertension)
ANTONSET$hypertension <- ifelse(ANTONSET$SysAvg >= 140 | ANTONSET$DiaAvg >= 90, "Yes", "No")

# Create a decision tree model
decision_tree_model <- rpart(hypertension ~ age + gender + BMI + Anxiety, data = ANTONSET, method = "cla

# We use a threshold of 140/90 mmHg for systolic (SysAvg) and diastolic (DiaAvg) blood pressure
```

Model construction: Search for the model that best fits the nature of the data using the training dataset. In our design, we will opt for a logistic regression model (analyzes the relationship between a binary dependent variable - whether a person has hypertension or not, or whether they experience psychotic episodes or not - and one or more independent variables - age, gender, BMI, physical activity, medication, etc.).

```
# Fit a logistic regression model 1
logistic_model1 <- glm(disease ~ age + gender + blood_pressure, data = data, family = binomial)

# Fit a logistic regression model 2
logistic_model2 <- glm(crisis ~ hypertension_threshold + climatology + weekly_physical_activity + medica

# View the estimated coefficients of the model
summary(logistic_model)
```

- Generation of a testing design based on the parameters used to create or search for the model in the previous step.
- Model evaluation - e.g., k-fold cross-validation technique [@james2013; @kuhn2008] to assess the performance of our logistic regression model that predicts the probability of developing hypertension based on variables such as mental health.
  - Identification of patterns in the data, analyzing and repeating the process until optimal parameters are found. In our hypertension and mental health project, we could use k-fold cross-validation to assess the performance of a logistic regression model, specifically the one predicting the probability of developing hypertension based on mental health variables such as depression and anxiety. We would divide the data into k subsets, train and evaluate the model k times using different subsets as training and validation sets.
  - Performance evaluation. The process is finalized or repeated until optimal parameters are found based on the evaluation results - e.g., calculating the average accuracy of the model on the k validation sets and making data-driven decisions based on time to obtain results and apparent accuracy.

Phase 5: Evaluation

Overall objective of the phase:

While the previous phase focused on evaluating technical aspects, this phase aims to evaluate whether the results meet the objectives established in Phase 1 - understanding the business. It involves a broader evaluative analysis.

Specific Objectives:

1. Evaluation of the results: Have we achieved the established objectives?
2. Review of the process: Review of the work carried out so far, summarizing findings, and reviewing procedures.
3. Determining the next steps: Based on the information obtained in the previous steps, decide whether to proceed with the next phases or iterate back to an earlier point in the model.

Product:

- Document and/or explanatory report of the results with an assessment of their usefulness.

Tasks:

- Evaluation of the results based on the objectives.
- Identification of the model's usefulness for making guided decisions.
- Determining whether to proceed to the next phase or consider going back to a previous phase for model adjustments.

Phase 6: Deployment

Overall objective of the phase:

Implementation of the data mining model in the organization.

Specific Objectives:

- Implementation of the model in the patient management environment without impacting system stability.
- Creation of an alert system guided by the model.

When risks of psychotic episodes, pathological behaviors, or hypertension are detected, healthcare staff will receive the corresponding notifications.

Product:

Implemented and documented data mining model.

Tasks:

- Preparation of the infrastructure for model implementation.
- Definition of the integration process with the organization's systems.
- Documenting the model and its usage for end users.
- Reviewing compliance with current legislation, especially regarding security, data protection, and management of computerized files, without disregarding other legal aspects.

Differences from other standards and the CRISP-DM model:

When selecting the foundational model for the data mining project, to avoid bias due to my data analyst background, I delved into various models and frameworks for data mining and the work of data scientists.

In this section, I intend to highlight the differences between the proposed data mining project and other models and conventions and justify the decision.

Typically, a data analyst works within a necessarily linear framework of 6 phases (*ask, prepare, process, analyze, share, and act*). It is not an iterative process but rather a process aimed at obtaining "use and discard" knowledge. In this framework, the analysis phase yields conclusions and predictions that serve to guide specific decisions. If the results are unsatisfactory, the analysis phase may be repeated or the work may be discarded, but it is never considered iterative, and the product is not a model with long-term utility. On the contrary, in a data mining model, the structure must be reviewed and improved while the model is in use.

The most commonly used framework based on the data lifecycle model is the one based on the CRISP-DM model. It consists of 5 phases. The second phase (data selection) encompasses the two data-related phases of CRISP-DM, "data understanding" and "data preparation". The remaining phases appear analogous. As a convention, the data science process based on the data lifecycle model consists of 6 phases: plan, capture, manage, analyze, archive, and destroy. The phase names are descriptive labels, but this base structure does not seem representative of the type of project I wanted to illustrate in this exercise. Therefore, it was necessary to study different models and proposals based on the course material to select the appropriate framework.

The CRISP-DM model differs from other data lifecycle models such as SEMMA and KDD in its iterative and project-oriented approach. It excels in iterating through the phases to improve the model, which is especially useful when a significant part of the work involves trial and error. SEMMA focuses on data analysis, while KDD focuses on knowledge extraction and the generation of predictive models. CRISP-DM is a more comprehensive and flexible model that can adapt to different projects and needs, and it aligns with the teaching proposal.

Similarly, agile models can be adapted and applied to data mining projects. In an agile approach, the data mining process is divided into smaller and more manageable iterations rather than complete planning from the beginning. This allows for greater flexibility and responsiveness to changes as the project progresses, which seems suitable for the proposed data mining project on the relationship between hypertension and mental health. It promotes greater collaboration and communication between the data mining team and end users, which is unnecessary in our case. Overall, it does not seem to align well with the teaching proposal.

Finally, other models such as Waterfall, Microsoft TDSP, and Domino Lifecycle can be adapted and applied to data mining projects. Waterfall is a software lifecycle model that focuses on the planning and execution of software projects in a linear manner. Although it includes some phases that overlap with CRISP-DM, such as requirements definition and implementation, it is not specifically designed for data mining. Microsoft TDSP (Team Data Science Process) is a framework specifically designed for data science projects and includes phases similar to those of CRISP-DM. However, it has a strong focus on teamwork and collaboration, making it more suitable for larger projects with a steep learning curve and implementation.

Domino Lifecycle focuses on collaboration and project management of data science projects through a cloud platform. It includes a set of tools and functions to facilitate collaboration and project management, as well as a set of best practices for data science. However, it requires the use of the platform, which can be costly.

Model	Main focus	Main phases	Advantages	Disaventages	
CRISP-DM	*Project-oriented*	***6 phases:*** *understanding the business, understanding the data, data preparation, modelling, assessment and deployment*	*structured and clear approach, adaptable to different projects and situations, widely used*	*Requires a detailed approach and a multidisciplinary team, not suitable for small projects.*	
SEMMA	*Project-oriented*	***5 phases:*** *sample, explore, modify, model, evaluate*	*Detailed approach to each phase, focusing on accurate modelling*	Can be too structured and rigid, not suitable for projects that require flexibility.	
KDD	*Process-oriented*	**9 phases**: *domain understanding, data selection, preprocessing, transformation, data mining, evaluation, presentation of results, implementation and maintenance.*		*Detailed and comprehensive approach, suitable for large and complex projects*	*Puede ser demasiado estructurado y rígido, requiere un equipo multidisciplinario y puede ser costoso*
Waterfall	*Linear approach*	***Sequential phases:*** *analysis, design, implementation, testing, maintenance*	*Clear and easy-to-follow approach, suitable for simple, well-defined projects*	*Not suitable for complex or changing projects, does not allow for feedback*	
Microsoft TDSP	Project Oriented	***5 phases:*** *project planning, data preparation, data exploration, model building, deployment*	*Structured and clear approach, with a focus on collaboration and communication between teams*	*Requires the use of specific Microsoft tools, not suitable for projects that do not use Microsoft technologies.*	

Model	Main focus	Main phases	Advantages	Disaventages
Domino Lifecycle	*Project-Oriented*	***7 phases:*** *problem identification, project definition, data preparation, data exploration, modelling, evaluation, deployment*	*Structured and clear approach, with a focus on collaboration and communication between teams*	*Requires use of 'Domino' platform, can be costly*

In summary, the proposed project revolves around the "Tasks" section in all its phases. Including tasks is a recommended practice in many data lifecycle models, ensuring that all important activities have been considered and are being carried out in the project. The use of tasks helps improve team efficiency and maintain a more organized project with better tracking, although in this case, it is an individual work. Additionally, examples and general explanations have been included in subheadings to make it more understandable. These aspects would not be present in a real project unless there is a need for specific clarifications for team members unfamiliar with the working methodology.

SPANISH

Proyecto de Minería de Datos: Relación entre Hipertensión y Psicopatologías

Fase 1: Entendiendo el negocio

Objetivo general de la fase: Definición del problema de investigación y establecimiento de los objetivos del proyecto.

Problema: Los responsables del área de salud mental de una empresa de geriátricos, detectan indicios que parecen apuntar a que la hipertensión arterial esta relacionada con un empeoramiento de los trastornos mentales y la aparición de brotes psicóticos y crisis.

Objetivos:

- *Refutación estadística*: Análisis de los datos históricos de la empresa para establecer la significancia estadística entre la hipertensión y los problemas de conducta y establecer correlación, causalidad o descartar la teoría.
- *Prevención y mejoras de la atención*: Mediadas por un modelo que permita la alerta temprana, estableciendo un modelo basado en umbrales de medidas fisiológicas e historiales clínicos, habilitando el establecimiento de medidas preventivas.

Producto: Documento base para la creación de un proyecto de *machine learning.* Este habilitaría implementar mejoras en el programa informático de gestión de salud de los pacientes, detectando parámetros de riesgo y lanzando alertas tempranas. El documento contendría un **desglose** de las variables relevantes al problema y **especificaría** los umbrales generales y específicos de tensión arterial susceptibles de afectar a cada paciente. ***Se trataria de un modelo predictivo.***

Tareas:

- *Definición* de la población de estudio - clientes alojados en algún asilo o instalación de la empresa-.
- Recopilación de *datos primarios* de las diferentes fuentes en poder de la empresa.
- Recopilación de *datos secundarios* relacionados con el problema, ya sean recopilados por otras instituciones o procedentes de investigaciones -*cabria en este proyecto propuesta evitar los datos de fuentes terciarias o cuyos datos provienen de terceros, que no han sido recopilados directamente por el publicador-*.
- *Identificación* de las variables relevantes para el análisis.

Criterios de éxito:

- Establecimiento de **correlación o causalidad** entre salud mental y estados de hipertensión arterial teniendo en cuenta que la hipertensión puede ser a la vez un factor que influya en el desarrollo de una enfermedad mental o ser un factor que exacerba la sintomatología una vez desarrollada una psicopatologia. En definitiva atenderemos al papel que juega la hipertensión y su relacion con las enfermedades mentales o neurológicas.
- Diseño de un **modelo base** para el desarrollo de un modulo o programa de aprendizaje automático o análogo desplegado en la empresa.

Fase 2: Entendiendo los datos

Objetivo general de la fase: Identificar y entender los datos necesarios para el desarrollo del proyecto de minería de datos.

Objetivos:

- Descarte de conjuntos de datos sin relación con el proyecto - *e.g: registros de compras, visitas de familiares, higiene, etc.*-
- Entendimiento de las características y estructura de los datos necesarios para el proyecto - *formatos y/o tipo de ficheros, localización, acceso, seguridad, numero de registros, atributos, etc.* -

Producto: Documento extensivo descriptivo de los datos disponibles - *calidad, fuente, antigüedad, relevancia, funciones en el proyecto, etc.* -

Tareas:

- *Análisis exploratorio de los datos* agenciado a extraer primeras conclusiones - *numero de registros, numero de variables, tipologia, distribución de los valores, errores (formato incorrecto, nulos, campos vacíos, espacios en blanco), los valores (puntuaciones extremas, tipo de formato inadecuado).* -

Fase 3: Preparación de los datos

Objetivo general de la fase: Preparación de los datos con base a los aspectos identificados en la fase 2 y adecuarlos para el trabajo de análisis.

Productos:

- Conjunto de datos limpios y transformado para posterior análisis.
- *changelog* o registro de cambios en los archivos de datos.

Tareas: *Reducción de la dimensionalidad*

- Reducción y filtrado de atributos y características - *e.g: Excluir las fechas que quedan fuera del estudios o del estudio o sujetos de menor edad que lo establecido en el proyecto.* -
- Eliminación de valores extremos que alteren la bondad del modelo - *e.g: hay valores extremadamente altos o bajos en la presión arterial que son improbables o incluso imposibles de medir y es necesario eliminarlos del conjunto de datos porque distorsionan la precisión del modelo.* -
  - Mediante análisis de componentes principales (PCA) o análisis de correlación canónica (CCA) para la reducción del conjunto de variables a pocos factores que explique la mayor parte de la variabilidad de los datos - *e.g: reducimos los registros de cada día a los valores representativos en tensión arterial* -.
  - Mediante matriz de correlaciones y eliminación iterativa de características (RFE) para la identificación de variables altamente relacionadas.
  - Otras técnicas de valores atípicos (análisis multivariante, análisis de cajas y bigotes).
- Condensación para reducir muestras
- Creación de nuevas variables para combinar o derivar variables existentes que pueda ser mas útil en el análisis - *e.g: tomar datos de los registros de ingesta de alimentos y cantidad de líquidos para crear una nueva variable 'Hidratación'*-.

Transformación:

- Normalización de datos
  - Escalamiento de los datos - *e.g: Cambiar metros por kilómetros recorridos en el registro de actividad física* -.
  - Normalización por el máximo - *e.g: Queremos normalizar la edad de los residentes, dividimos la edad de cada residente por la edad máxima, derivando en una escala de entre 0 y 1* -
  - Normalización por la diferencia - *e.g: Queremos normalizar la dosis diaria de antipsicóticos (Risperidona)*

    `Dosis normalizada= (Dosis-Dosis máxima)/(Dosis máxima- Dosis mínima)= Valor entre 0 y 1.-`

  - Normalización por la desviación estándar - *e.g: la variable "dosis de Risperidona"* -.
    1. calculamos la desviación estándar de la variable:

       ```
       dosis_sd <- sd(datos$dosis_Risperidona)
       ```

    2. normalizamos la variable dividiéndola por la desviación estándar:

       ```
       datos$dosis_Risperidona_norm <- datos$dosis_Risperidona/dosis_sd
       ```

       Y obtendríamos la nueva variable "dosis de Risperidona normalizada".
- Codificación de variables categóricas, asignando valores numéricos para poder acometer análisis estadístico -e.g: Para nuestro modelos, transformamos actividad física (Sedentario, moderadamente activo, muy activo) en valores numéricos tales que 1,2 y 3 respectivamente-.

Limpieza de los datos:

- *Identificación de sesgos* - ideológicos, de genero, etc.-, en los datos y/o en las muestras -sesgos del observador, de interpretación, de confirmación -.
  - Ejemplo sesgo ideológico: *Un investigador tiene una posición política muy definida y quiere demostrar que la hipertensión está fuertemente relacionada con la falta de acceso a atención médica, seleccionando datos o interpretarlos de manera sesgada para respaldar su hipótesis, incluso ignorando otras variables importantes.*
  - Ejemplo de sesgo de genero: un investigador puede estar más interesado en la relación entre la enfermedad y los hombres, y consecuentemente, centrar la atención en este grupo, ignorando factores importantes en la hipertensión en las mujeres, como la menopausia, el embarazo o la anticoncepción hormonal.
  - Ejemplo de sesgo del observador: *Un investigador esta interesado en relacionar el estrés como factor de riesgo para la hipertensión porque hay casos en su familia. Consecuentemente observa principalmente los casos donde el estrés se presenta con la hipertensión.*
  - Ejemplo de sesgo de interpretación: *Un investigador puede encontrar una correlación entre la hipertensión y la edad, pero podría interpretar erróneamente que la hipertensión es causada por la edad, en lugar de que la edad es un factor de riesgo para la hipertensión.*
  - Ejemplo de sesgos de confirmación: *un investigador podría tener una hipótesis previa sobre la relación entre la enfermedad y el consumo de sal, y buscar datos o análisis que respalden su hipótesis. Esto lleva a ignorar otros factores que podrían contribuir a la hipertensión, como la genética o el estilo de vida en general.*
- *Garantizado* de la fiabilidad, originalidad, comprensibilidad, actualidad o validez y las fuentes de los datos con los que se va a operar.
- *Limpieza* propiamente dicha de los datos atendiendo a aspectos de los datos como:
  - Datos incompletos - *e.g: faltan registros de toma algunas tomas de presión arterial algunos días. No podemos inferir mediante regresión lineal dada la naturaleza del proyecto y decidimos eliminar los días que no están completos-*
  - Datos incorrectos - *e.g: Una fecha de 1705 en el campo 'fecha' respecto a una fecha de nacimiento debe tratarse de un error -*
  - Integridad de los datos o carencia de precisión, completidad, consistencia y/o credibilidad - *e.g: Diferentes formatos de identificador como clave primaria, medidas de tensión arterial de dos cifras, etc. -*
  - Datos duplicados - *e.g: Una misma instancia con el mismo nombre de usuario se repite en 3 registros, por lo que decidimos integrarlos o fusionarlos -*
  - Eliminación de espacios en blanco - *e.g: dos espacios para los nombres compuestos en vez de uno -*
  - Corrección ortográfica y de errores de escritura.
  - Asegurado de la compatibilidad de los juegos de datos para poder trabajar con claves foráneas y poder fusionar datos.
  - Formatos incorrectos - *e.g: Formato de texto para fechas en vez del formato de fecha adecuado a la estructura con la que estamos trabajando. -*

Fase 4: Modelado

Objetivo general de la fase: Construcción de un modelo de minería de datos y evaluación del rendimiento.

Producto:

- Modelo de minería de datos sobre la hipertensión y psicopatologías en población anciana, tanto la capacidad de predecir problemas conductuales ante la concurrencia de diversos parámetros como la capacidad de anticipar problemas de hipertensión ante determinados parámetros y valores de pruebas psicofísicas.
- Evaluación del modelo

Tareas:

- Selección de las técnicas de modelado. Este punto concreto gira en torno a la selección de algoritmos *- Usaremos algoritmo de árboles de decisión que permitan predecir la probabilidad de padecer hipertensión en virtud de variables como la edad, genero, IMC, ansiedad o depresión y el riesgo de sufrir brotes psicóticos en virtud de la dosis de medicación, condiciones atmosféricas, genero, edad y determinados umbrales de hipertensión. Los algoritmos de árbol son el conjunto de decisiones que se basan en la construcción de un árbol donde cada nodo representa una variable de decisión y cada rama representa una posible respuesta a esa variable. **El objetivo es que sea capaz de clasificar de manera precisa nuevas observaciones.** Para ello, se busca la variable de decisión que mejor separa las observaciones en función de su clase - la variable que permite hacer la clasificación más precisa -. Se evalúan diferentes variables de decisión y se selecciona la que proporciona el mayor beneficio en términos de precisión

  Ejemplo practico en R:

  Queremos predecir si una persona tiene hipertensión o no en función de su edad, género, índice de masa corporal (IMC) y nivel de ansiedad:

```
# Carga de la librería
library(rpart)

# Carga del conjunto de datos ANTONSET que deriva de las fases anteriores
data(ANTONSET)

# Creacion de una columna para la variable objetivo (hipertensión)
ANTONSET$hipertension <- ifelse(ANTONSET$SysAvg >= 140 | ANTONSET$DiaAvg >= 90, "Si", "No")

# Creacion de modelo de árbol de decisión
modelo_arbol <- rpart(hipertension ~ edad + genero + BMI + Anxiety, data = ANTONSET, method = "clas
# utilizamos un umbral de presión arterial sistólica (SysAvg) y diastólica (DiaAvg) de 140/90 mmHg.
```

- Construcción del modelo: Búsqueda del modelo que mejor responde a la naturaleza de los datos utilizando el conjunto de datos de entrenamiento - *en nuestro diseño nos decantaremos por un modelo de regresión logística (analiza la relación entre una variable binaria dependiente - si una persona tiene o no hipertensión o si sufre crisis psicóticas o no- y una o más variables independientes - edad, género, IMC, , actividad física, medicación, etc.)* Hosmer Jr, Lemeshow, and Sturdivant (2013)

```
# ajustar un modelo de regresión logística 1
modelo_logistico1 <- glm(enfermedad ~ edad + genero + presion_arterial, data = datos, family = binc
# ajustar un modelo de regresión logística 2
modelo_logistico2 <- glm(crisis ~ umbral_hipertension + climatologia + actividad_fisica_semanal + r

# ver los coeficientes estimados del modelo
summary(modelo_logistico)
```

- Generación de un diseño de pruebas conforme a los parámetros que hayamos usado para crear o buscar el modelo en el punto anterior.
- Evaluación del modelo - *e.g: técnica de validación cruzada k-fold* (James et al. 2013; Kuhn 2008) *para evaluar el rendimiento de nuestro modelo de regresión logística que predice la probabilidad de desarrollar hipertensión en función de variables como la salud mental.-*
  - Identificación de patrones en los datos, se analizan y se repite el proceso hasta encontrar parámetros óptimos. - *En nuestro proyecto de hipertensión y problemas de salud mental, podríamos utilizamos la validación cruzada k-fold para evaluar el rendimiento de un de los modelos de regresión logística, en este caso el que predice la probabilidad de desarrollar hipertensión en función de variables de salud mental como la depresión y la ansiedad. Dividiríamos los datos en k subconjuntos, entrenaríamos y evaluaríamos el modelo k veces utilizando diferentes subconjuntos como conjuntos de entrenamiento y validación.*
  - Evaluación del rendimiento. Se procede a finalizar o repetir el proceso hasta encontrar los parámetros óptimos consecuentemente a los resultados de la evaluación - *e.g: calculo de la precisión promedio del modelo en los k conjuntos de validación y toma de decisiones orientada por los datos, tiempo en obtener los resultados y aparente precisión.*

Fase 5: Evaluación

Objetivo general de la fase: Mientras en la fase anterior nos centramos en evaluar los aspectos tecnicos, en esta fase evaluaremos si los resultados cumplen con los objetivos establecidos en la fase 1 - *entendimiento del negocio* -. Es por tanto un análisis evaluativo mas amplio "What Is CRISP DM? - Data Science Process Alliance" (n.d.)

Objetivos (específicos)

1. Evaluación de los resultados: ¿Hemos alcanzado los objetivos establecidos?
2. Revisión del proceso: Revisión del trabajo llevado a cabo hasta el momento, sumarización de hallazgos y revisión de procedimientos
3. Determinación de los siguientes pasos: Con la información obtenida en los pasos anteriores, decidir si seguir adelante con las fases o iterar a un punto anterior del modelo.

Producto:

- Documento y/o memoria explicativa de los resultados con valoración de la utilidad

Tareas:

- *Evaluación* de los resultados en virtud de los objetivos.
- *Identificación* de la utilidad del modelo para poder tomar decisiones guiadas.
- *Determinación* de de la conveniencia de avanzar o no a la siguiente fase o establecer si seria necesario retrotraerse a una fase anterior para realizar reajustes al modelo.

Fase 6: Despliegue

Objetivo general de la fase: Implementación del modelo de minería de datos en la organización.

Objetivos específicos:

- *Implementación* del modelo en el entorno de gestión de pacientes sin que repercuta en la estabilidad del sistema
- *Creación* de un sistema de alertas guiado por el modelo. Ante la detecccion de riesgos de brotes psicóticos, problemas de comportamientos patológicos o de hipertensión, el personal sanitario recibirá las correspondientes notificaciones.

Producto:

- Modelo de minería de datos implementado y documentado

Tareas:

- Preparación de la infraestructura para la implementación del modelo.
- Definición del proceso de integración con los sistemas de la organización.
- Documentar el modelo y su uso para usuarios finales.
- Revisión del cumplimiento de la legislación vigente, especialmente en lo referido a la seguridad, protección de datos y la gestión de ficheros informatizados, sin descartar otros aspectos legales.

Diferencias del modelo propuesto respecto a otros estándares y el modelo CRISP DM A la hora de seleccionar el modelo de base para el proyecto de minería de datos, para evitar un sesgo debido a la formación como analista de datos, he abordado un periodo de interiorizacion de las diferentes modelos o marcos para la minería de datos y el trabajo de científico de datos. Pretendo en este apartado resaltar las diferencias entre la propuesta de proyecto de minería de datos respecto a otros modelos y convenciones y justificar la decisión.

Por lo general, un analista de datos trabaja en un marco **necesariamente lineal** de 6 fases (*ask, prepare, process, analyze, share y act*). No se trata de un proceso iterativo, sino mas bien de un proceso guiado a la obtención de conocimiento de "usar y tirar". En este, la fase de análisis, se extraen conclusiones, predicciones que sirven para orientar decisiones puntuales. En el caso de que los resultados no sean satisfactorios, puede repetirse la fase de análisis o desechar el trabajo, pero, en ningún caso puede considerarse iterativo ni el producto es un modelo con una utilidad dilatada en el tiempo. al contrario, en un modelo de minería de datos, la estructura debe revisarse y mejorarse mientras el modelo esta vigente.

El marco de trabajo basado el modelo del ciclo de vida de los datos mas usual es el basado en el modelo CRISP DM "What Is CRISP DM? - Data Science Process Alliance" (n.d.) Este consta de 5 fases. La segunda fase (selección de datos) englobaría las dos fases de trabajo con datos de CRISP DM, "*data understanding*" y "*data preparation*". El resto de fases parecen análogas. Como convención, el proceso de ciencia de datos basado en el marco del ciclo de vida de los datos consta de 6 fases: *plan, capture, manage, analyze, archive y destroy.* Los nombres de las fases son etiquetas descriptivas, pero esta estructura base no parece ser representativa del tipo del proyecto que quería ilustrar en este ejercicio. Por todo ello, ha sido necesario con base al material de la asignatura, estudiar los diferentes marcos y propuestas a la hora de seleccionar el marco de trabajo adecuado.

El modelo CRISP-DM se diferencia de otros modelos de ciclo de vida de los datos como SEMMA y KDD en su enfoque iterativo y orientado a proyectos. Destaca en la iteración de las fases para mejorar el modelo. Esto es especialmente útil cuando una parte del trabajo gira en torno al "prueba y error". SEMMA se enfoca

en el análisis de los datos, mientras que KDD se enfoca en la extracción de conocimiento y en la generación de modelos predictivos. CRISP-DM es un modelo más completo y flexible que permite adaptarse a diferentes proyectos y necesidades y además, entiendo que la propuesta de la docencia gira en torno a él.

Al igual que CRISP-DM, los modelos de ciclo de vida de datos como *Waterfall, Microsoft TDSP y Domino Lifecicle* también se enfocan en el proceso de extracción de conocimiento a partir de los datos, aunque las diferencias son importantes. Por un lado, *Waterfall* es un modelo de ciclo de vida de software que se centra en la planificación y ejecución de proyectos de software de manera **linea**l. Aunque incluye algunas fases que se solapan con CRISP-DM, como la definición de requisitos y la implementación, no está diseñado específicamente para la minería de datos. Aunque *Microsoft TDSP (Team Data Science Process)* es un marco de trabajo diseñado específicamente para proyectos de ciencia de datos e incluye fases similares a las de CRISP-DM, tiene también una fuerte orientación hacia el trabajo en equipo y la colaboración, lo que lo hace adecuado para proyectos más grandes y una elevada curva de aprendizaje e implementación.

Por su parte, *Domino Lifecicle* se centra en la colaboración y la gestión de proyectos de ciencia de datos a través de una plataforma en la nube. Incluye una serie de herramientas y funciones para facilitar la colaboración y la gestión de proyectos, así como una serie de mejores prácticas para la ciencia de datos. Requiere el uso de la plataforma y por tanto, tiene un coste elevado.

Finalmente, los modelos ágiles se pueden adaptar y aplicar a proyectos de minería de datos. En un enfoque ágil, el proceso de minería de datos se divide en iteraciones más pequeñas y manejables en lugar de una planificación completa desde el principio. Esto permite una mayor flexibilidad y capacidad de respuesta a los cambios a medida que se avanza en el proyecto, lo que parece adecuado a la propuesta de proyecto de minería de datos sobre la relación de la hipertensión aquí articulada. Promueve una mayor colaboración y comunicación entre el equipo de minería de datos y los usuarios finales, lo que es innecesario en nuestro caso. Finalmente y en general, no parece muy compatible con la propuesta docente

Modelo	**Enfoque principal**	**Fases principales**	**Ventajas**	**Desventajas**
CRISP-DM	*Orientado a proyectos*	***6 fases**: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue*	*Enfoque estructurado y claro, adaptable a diferentes proyectos y situaciones, ampliamente utilizado*	*Requiere un enfoque detallado y un equipo multidisciplinario, no es adecuado para proyectos pequeños*
SEMMA	*Orientado a procesos*	***5 fases:** muestra, explora, modifica, modelo y evalúa*	*Enfoque detallado para cada fase, enfocado en la creación de modelos precisos*	*Puede ser demasiado estructurado y rígido, no adecuado para proyectos que requieren flexibilidad*

Modelo	Enfoque principal	Fases principales	Ventajas	Desventajas
KDD	*Orientado a procesos*	***9 fases:*** *comprensión del dominio, selección de datos, preprocesamiento, transformación, minería de datos, evaluación, presentación de resultados, implementación y mantenimiento*	*Enfoque detallado y completo, adecuado para proyectos grandes y complejos*	*Puede ser demasiado estructurado y rígido, requiere un equipo multidisciplinario y puede ser costoso*
Waterfall	*Enfoque lineal*	***Fases secuenciales:*** *análisis, diseño, implementación, pruebas, mantenimiento*	*Enfoque claro y fácil de seguir, adecuado para proyectos simples y bien definidos*	*No es adecuado para proyectos complejos o cambiantes, no permite la retroalimentación*
Microsoft TDSP	*Orientado a proyectos*	***5 fases:*** *planificación del proyecto, preparación de los datos, exploración de los datos, construcción del modelo, despliegue*	*Enfoque estructurado y claro, enfocado en la colaboración y la comunicación entre los equipos*	*Requiere el uso de herramientas específicas de Microsoft, no es adecuado para proyectos que no utilicen tecnologías de Microsoft*
Domino Lifecycle	*Orientado a proyectos*	***7 fases:*** *identificación del problema, definición del proyecto, preparación de los datos, exploración de los datos, modelado, evaluación, despliegue*	*Enfoque estructurado y claro, enfocado en la colaboración y la comunicación entre los equipos*	*Requiere el uso de la plataforma 'Domino', puede ser costoso*

Finalmente, la propuesta aquí articulada gira en torno a un epígrafe 'tareas' en todas sus fases. La inclusión de las tareas es una práctica recomendada en muchos modelos de ciclo de vida de la ciencia de datos, y que ayuda a garantizar que todas las actividades importantes se hayan considerado y se estén llevando a cabo en el proyecto Ahmed et al. (2021) El uso de tareas ayuda a mejorar la eficiencia del equipo y a mantener un proyecto más organizado y con un mejor seguimiento, aunque en este caso se trata de un trabajo individual. Asimismo en subapartado se han incluido ejemplos y explicaciones generalistas globales sobre la naturaleza de cada sección para hacerlo mas comprensible. Estos aspectos no estarían presentes en un trabajo real salvo la necesidad de aclaraciones puntuales para miembros del equipo no familiarizados con la metodología de trabajo.

BIBLIOGRAPHY

Ahmed, Syed Thouheed, Syed Muzamil Basha, Sajeev Ram Arumugam, and Kiran Kumari Patil. 2021. *Big Data Analytics and Cloud Computing: A Beginner's Guide*. MileStone Research Publications.

Hosmer Jr, David W, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied Logistic Regression*. Vol. 398. John Wiley & Sons.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 112. Springer.

Kuhn, Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28: 1–26.

"What Is CRISP DM? - Data Science Process Alliance." n.d. https://www.datascience-pm.com/crisp-dm-2/.