En luz de las nuevas y populares tendencias en el mundo de la estadística computacional, llamada en ocasiones aprendizaje estadístico u aprendizaje de máquina;[1] este trabajo plantea como objetivo: estudiar, explicar e implementar un modelo de clasificación supervisada con base en un modelo *probit* al que se le añade un componente no lineal de bases aditivas. Asimismo, se busca desarrollar un algoritmo asociado de aprendizaje para la estimación de parámetros con base en el paradigma bayesiano de inferencia.[2]

El modelo, buscará hacer inferencia sobre una base de datos y *aprender* sobre los patrones subyacentes que puedan contener para posteriormente, predecir el resultado de variables de respuesta. Este tipo de modelos, han resultado ser de enorme efectividad en ámbitos tan diversos, como lo son la medicina y las finanzas. Bajo esta óptica, se busca que el modelo sea práctico e útil, sin perder de vista el componente teórico subyacente. Por lo tanto, se busca explicar con el mayor detalle, cada componente del modelo para que este no sea tratado como una caja negra computarizada.

Los modelos probit son un tipo de regresión, que busca la clasificación de variables de respuesta y_i binarias (éxito o fracaso, positivo o negativo, etc).[3] Esta predicción, depende de información contenida en las covariables $\mathbf{x}_i$ para cada una de las obser-

1. *machine learning (ML)*
2. Es común, hacer una distinción entre el aprendizaje estadístico y el aprendizaje de máquina pues, mientras que los modelos son los mismos, difieren en perspectiva. Mientras que el aprendizaje estadistico presta mayor atención al aspecto inferencial e interpretación, el aprendizaje de máquina coloca mayor énfasis en la implementación computacional y los resultados.
3. Es usual en la literatura, hablar de *clasificadores* cuando las respuestas son categorías (codificadas en variables discretas) y *regresiones* cuando las variables de respuestas son continuas.

vaciones $i = 1, \ldots, n$. Sin embargo, la relación entre y_i con $\mathbf{x}_i$ puede depender de estructuras complejas que no son necesariamente lineales, esto lleva a que la predicción de las respuestas y_i con base en $\mathbf{x}_i$ sea difícil. Para sobrepasar esto, al modelo se le agrega un componente no lineal en las covariables que permite discernir estos patrones. Como se verá en el trabajo, el modelo induce fronteras no lineales de clasificación en el espacio donde $\mathbf{x}_i$ tome valore. En la figura 1, se tiene un ejemplo gráfico de tipo de clasificación que lleva a cabo el modelo. Se tienen observaciones del grupo azul y del grupo rojo con una clara separación no lineal en las covariables x_1 y x_2. El proceso de aprendizaje busca *entrenar*, bajo el paradigma bayesiano, a una función η que logre separar este espacio de la mejor forma posible. Esta separación, induce una clasificación binaria (0 y 1 correspondiendo a rojo y azul respectivamente) a través de la función de distribución normal Φ. Con un modelo de clasificación cuya frontera fuera lineal en covariables, llevar a cabo esta clasificación sería imposible.

Para llevar a cabo la construcción del modelo, se comienza con una discusión teórica en el capítulo **??**. Primeramente se estudian los modelos lineales generalizados (GLM), específicamente los modelos probit. Los GLM como su nombre lo indica, generalizan las regresiones tradicionales donde la respuesta y_i es escalar ($y_i \in \mathbb{R}$) a regresiones donde la respuesta puede ser discreta o restringida a cierto dominio (**maccullagh1989generalized**). No obstante, los GLM siguen siendo lineales en las covariables pero se pueden flexibilizar usando diferentes ideas. Entre ellas, los modelos aditivos generalizado (GAM) presentadas en **hastie1986generalized**. En estos modelos, la flexibilización se logra transformando a las covariables $\mathbf{x}_i$, mediante la función de predicción η, usando métodos no paramétricos con base en

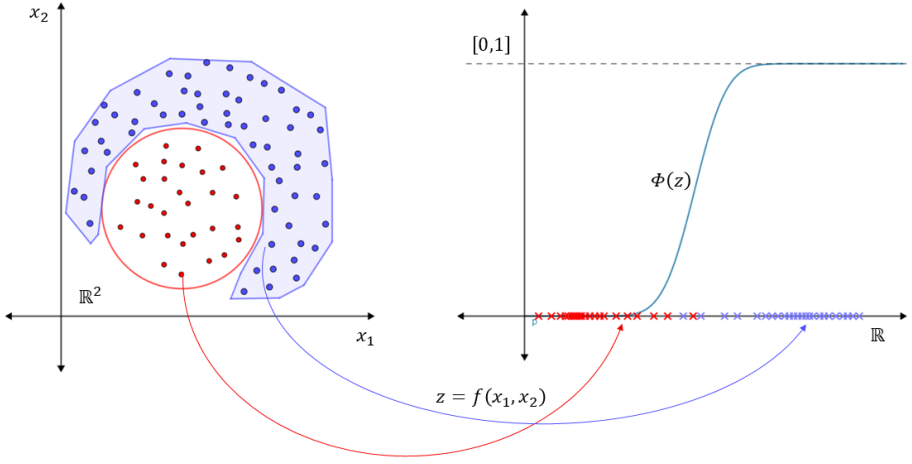

Figura 1: Diagrama explicativo de un modelo de clasificación probit no lineal

suavizadores. Para este trabajo, se toman esas ideas y se combinan con las de **mallik1998automatic** en las que se opta por darle una forma funcional concreta a η, correspondiente a una expansión de bases funcionales, particularmente, en polinomios por partes de continuidad y grado arbitrarios. La expansión resultante, tiene la peculiaridad que conectan muchas disciplinas y ramas de las matemáticas que han sido de mucha utilidad no sólo en el campo de la estadística. A lo largo del capítulo, se verá que con principios presentados, se abren las posibilidades en cuanto a modelos y datos sobre los que se pueden hacer regresiones o clasificaciones.

Desarrollado una vez el modelo, el capítulo **??** se concentra en su implementación bajo el paradigma bayesiano de aprendizaje. Por lo tanto, se hace una breve introduc-

ción al paradigma bayesiano de la estadística, en particular al apredizaje bayesiano en un contexto de regresión. Este paradigma, responde a que, bajo las ideas de **albert1993bayesian**, se puede plantear un algoritmo asociado al modelo de forma tal que se induce un sampleo de Gibbs. La implementación final, se realiza en el paquete computacional para el lenguaje abierto de programación estadística `R`.[4]

En el capítulo **??**, el modelo se prueba y se valida contra una serie de bases de datos. Primeramente, se hace una breve discusión sobre como evaluar la efectividad y precisión de un modelo como el presentado en este trabajo. Posteriormente, se ejecuta el algoritmo de aprendizaje contra cinco bases de datos simulados con dos covariables ($\mathbf{x}_i \in \mathbb{R}^2$). Estas pruebas preliminares, sirven para demostrar las capacidades predictivas del modelo y sobre todo, para hacer más concretas las matemáticas subyacentes, además de poder visualizar las diferentes fronteras flexibles obtenidas por el modelo. Asimismo, en este capítulo se discute la convergencia de las cadenas obtenidas por el muestreador de Gibbs, uno de los puntos cruciales al trabajar con modelos de este tipo. Para cerrar el capítulo, se replica un escenario real de análisis y modelado, usando una base de datos médicos de cáncer de donde se obtienen buenos resultados.

Finalmente, se cierra la discusión en el Capítulo **??** donde se revisan consideraciones finales y limitantes del modelo, sin embargo, se abre una discusión a posibles extensiones para mejorarlo. Posteriormente, se da un rápido vistazo a modelos relativamente más modernos los cuales han sido capaces de proezas computacionales

4. El desarrollo y explicación del paquete de cómputo se detalla en el Apéndice **??**. El paquete se puede descargar libremente de: `https://github.com/PaoloLuciano/bpwpm`

que se creían imposibles hace algunas décadas. No obstante, se verá que muchos de estos modelos más complejos, son generalizaciones de modelos clásicos y extensiones del presentado en este trabajo.