Diagnóstico de enfermedades cardíacas con los algoritmos supervisados Naives Bayesian

Carlos Daniel González Cedillo[i]

Resumen

Las enfermedades cardíacas son la principal causa de muerte en la actualidad. Este paper contrasta la performance de los diferentes algoritmos supervisados de Machine Learning, que tienen aplicaciones en el área de la medicina, con los algoritmos supervisados Naives Bayes para ayudar a clasificar pacientes propensos a sufrir enfermedades cardíacas. Como fuente de datos se usan 303 instancias de pacientes con diferentes características que fueron analizados al procesar los datos con los respectivos algoritmos.

Los resultados con el algoritmo de Naives Bayes son prometedores, obteniendo una precisión del 86,81 %, usando la fuente de datos mencionada. Esta familia de algoritmos tiene un mejor rendimiento comparado con otros algoritmos de Machine Learning como neural networks, arrojando resultados más precisas que las esperadas de médicos humanos.

Abstract

Heart disease is the leading cause of the death in the present. This paper contrasts the performance between the different supervised algorithms of Machine Learning, applied in medicine field, with the Naive Bayes supervised algorithms to help classify patients prone to heart disease. As data source, 303 instances of patients with different characteristics were used and analized when the data was proccessed by the respective algorithms.

The results with the Naives Bayes algorithm are promising, obtaining an accuracy of 86.81 % using the mentioned data source. This family of algorithm has a better performance compared to other Machine Learning algorithms such as Neural Networks, obtaining more precise results than those expected from humans doctors.

[i] Estudiante avanzado de la Universidad de Palermo. carlos.dgonzalezc46@gmail.com

I. Introducción

Con los avances tecnológicos que experimentamos día a día, lo que parecía imposible hace unos años, ahora se encuentra a nuestro alcance. En el área de la salud, la tecnología ha tenido un gran impacto y la realidad es que se han conseguido grandes progresos a lo largo de las últimas décadas.

Hoy en día hay enfermedades que pueden ser prevenidas llevando un tratamiento adecuado e incluso llevando un mejor estilo de vida, pero el problema es saber identificar cuando una persona es propensa a sufrir dichas enfermedades para que así puedan ser tratadas a tiempo.

Las enfermedades de corazón son una de las enfermedades que más registros tienen y que pueden ser provocadas por diferentes factores. Estas enfermedades son más comunes de lo que se cree y muchas veces se hacen presentes de manera inesperada.[1]

Hay numerosos factores por los cuales las personas pueden tener problemas cardíacos, entre ellos tenemos: los factores hereditarios, como presión alta, historial familiar de enfermedades cardíacas, los factores congénitos, como el diábetes o los factores desarrollados por el estilo de vida que se lleva (fumar, mala alimentación).[2]

Existen señales tempranas que pueden ayudar a identificar dichas enfermedades en los pacientes y numerosos estudios y análisis médicos que han ayudado a detectar estas enfermedades de forma anticipada.[3]

En medicina, existen implementaciones de algoritmos predictivos con técnicas de Machine Learning y Data Mining en diferentes campos.[4] Hay investigaciones e incluso varias implementaciones para detectar tumores cancerígenos, clasificar el cáncer de acuerdo a diferentes categorías de diagnósticos, incluso para detectar fallas del corazón.[5] [6] Estas técnicas son bastante usadas para identificar patrones que ayuden a detectar enfermedades y son herramientas que en el campo de la medicina son utilizadas cada vez con mayor frecuencia. Existen diferentes algoritmos que son utilizados para sistemas predictivos y cada uno arroja resultados con diferentes grados de precisión dependiendo de los datos utilizados y de los resultados esperados.[7]

Existen 3 tipos de algoritmos de aprendizaje dentro del Machine Learning: Supervisados, no supervisados y por refuerzo. Dependiendo de las necesidades del problema a resolver, el ambiente en el que se van a desenvolver y los factores que afectarán la toma de decisiones, cada uno de estos tipos de aprendizajes nos ofrecen un mejor enfoque con respecto a los demás.[8]

Para estudios médicos, se han realizado comparaciones entre diferentes algoritmos. Entre ellos encontramos: constructores de árboles de decisión (Assistant- R, Assistant-L, LFC),[9] dos variantes de clasificadores Bayesianos (Clasificador Naives Bayes y semi- Naives Bayes),[10] [11] algoritmo de k vecinos más cercanos,[12] entre otros.[13]

El objetivo es desarrollar un sistema predictivo para determinar pacientes que

puedan ser propensos a sufrir enfermedades cardíacas, a través del uso de dos algoritmos supervisados, Naive Bayesian y Semi-Naive Bayesian y comparar estos resultados con las técnicas de Machine learning ya existentes. La idea es de entrenar el sistema con datos reales de pacientes donde se tomen en cuenta datos médicos (presián, diábetes, etc.) y datos personales (edad, sexo) para poder así determinar, de acuerdo a ciertas características, que personas pueden ser propensas a sufrir algún tipo de enfermedad del corazán. El sistema además de predecir cuáles pacientes son más propensos a sufrir enfermedades cardíacas, determina que tipo de deficiencias son las que los afectarían.

Este trabajo está estructurado de la siguiente manera: hay una primera sección que habla acerca de la enfermedades cardíacas (tipos, síntomas, causas, factores de riesgo), luego en la sección III se habla del Machine Learning y los tipos de aprendizajes que existen. En la sección IV se nombran los algoritmos que se aplican dentro del Data Mining, cuáles son los más utilizados y se explica en detalle el funcionamiento del Naives Bayesian y Semi-Naives Bayesian. Luego, hay una sección donde se describe la fuente de datos que se utiliza para implementar el sistema inicial. En la sección VI se realiza un análisis de los resultados obtenidos en la implementación del algoritmo y para finalizar se tiene un sección para la conclusión del trabajo y para comentar trabajos futuros a realizar.

II. Enfermedades cardíacas

Las enfermedades cardíacas describen una gama de enfermedades que afectan al corazón.[14] Comprenden enfermedades de los vasos sanguíneos, enfermedades de las arterias coronarias, problemas con el ritmo cardíaco (arritmias) y defectos cardíacos con los que has nacido (defectos cardíacos congénitos), entre otros.

Estas enfermedades se manifiestan de diferentes maneras y si bien algunas pueden manifestarse de forma inesperada, muchas otras pueden ser tratadas si son detectadas a tiempo.

En esta sección se estudia los tipos de enfermedades cardíacas, cuáles son los síntomas y las principales causas, y cuáles son los factores de riesgos más comunes para estas enfermedades.

II-A. Tipos

Existen diferentes tipos de enfermedades cardíacas que afectan diferentes partes del corazón y que son producidas por diversos factores.[15]

En esta parte mencionamos los diferentes tipos de enfermedades cardíacas que existen.

- Enfermedades Congénitas

- Arritmias
- Enfermedades de la arteria coronaria
- Miocardiopatía dilatada
- Infarto del Miocardio
- Insuficiencia Cardíaca

II-B. Factores de Riesgo

Las enfermedades relacionadas con el corazón también tienden a ser más probables en personas que cumplen ciertas características o que han estado expuesta a ciertas condiciones. Para este caso mostramos los factores que aumentan estas probabilidades de padecer alguna enfermedad cardíaca.[16]

- Edad. El envejecimiento aumenta el riesgo de que las arterias se dañen y se estrechen, y de que el muésculo cardíaco se debilite o engrose.
- Sexo. En general, los hombres corren mayor riesgo de padecer enfermedades cardíacas. Sin embargo, el riesgo para las mujeres aumenta después de la menopausia.
- Antecedentes familiares. Los antecedentes familiares de enfermedades cardíacas aumentan tu riesgo de padecer enfermedad de las arterias coronarias, especialmente, si uno de tus padres la desarrolló a temprana edad (antes de los 55 años para un familiar hombre, como tu hermano o tu padre, y antes de los 65 años para una familiar mujer, como tu madre o hermana).
- Fumar. La nicotina contrae los vasos sanguíneos, y el monóxido de carbono puede dar su revestimiento interno, lo que los vuelve más propensos a la aterosclerosis. Los ataques cardíacos son más frecuentes en fumadores que en no fumadores.
- Algunos medicamentos de quimioterapia y radioterapia contra el cáncer. Tal vez aumente el riesgo de padecer enfermedades cardiovasculares con algunos medicamentos de quimioterapia y las radioterapias.
- Mala alimentación. Una dieta con alto contenido de grasas, sal, azúcar y colesterol puede contribuir a causar la enfermedad cardíaca.
- Presión arterial alta. La presión arterial alta no controlada puede producir el endurecimiento y el engrosamiento de las arterias, lo que estrecha los vasos por los que circula la sangre.
- Niveles altos de colesterol en sangre. Los niveles altos de colesterol en sangre pueden aumentar el riesgo de que se formen placas y de aterosclerosis. Diabetes. La diabetes aumenta el riesgo de enfermedades cardíacas. Ambas afecciones comparten factores de riesgo similares, como obesidad y presión arterial alta.
- Obesidad. El exceso de peso normalmente empeora otros factores de riesgo.
- Falta de actividad física. La falta de ejercicio también está asociada con muchas formas de enfermedad cardíaca y con algunos de sus otros factores de riesgo.

- Estrés. El estrés sin tratar puede dañar las arterias y empeorar otros factores de riesgo de enfermedades cardíacas.
- Higiene deficiente. No lavarte las manos de forma regular y no generar otros hábitos que pueden ayudarte a prevenir las infecciones víricas o bacterianas puede ponerte en riesgo de contraer infecciones cardíacas, especialmente, si ya tienes una afección cardíaca no diagnosticada. La higiene dental deficiente también puede contribuir a las enfermedades cardíacas.

III. Machine learning

En Machine Learning pueden existir diferentes enfoques dependiendo de los resultados que se quieran obtener y de los datos disponibles para obtener dichos resultados. Aprendizaje supervisado, donde las instancias son dadas con etiquetas conocidas, es decir, con resultados esperados para cada entrada, el aprendizaje no supervisado, donde las instancias no vienen etiquetadas, se van creando y agrupando las instancias de acuerdo a la similitud de los resultados. Por último tenemos el aprendizaje por refuerzo, en este caso el sistema se expone a un ambiente donde se entrena a través del método de "prueba y error", el entrenamiento busca realizar una decisión mucho más específica. Dependiendo de la problemática que se tenga, se querrá implementar alguno de estos tres tipos de aprendizaje.[8]

También tenemos técnicas o métodos de conjuntos, que buscan combinar múltiples clasificadores de alguna manera. Se busca que las decisiones individuales de cada uno de estos clasificadores sean combinadas para clasificar nuevos registros.[17]

Una condición necesaria y suficiente para aplicar este método de conjuntos y que el conjunto sea más preciso que cualquiera de sus clasificadores individuales es cuando los clasificadores son precisos y diversos.[17] Dos clasificadores son diversos cuando presentan errores distintos.

Para el éxito de un sistema predictivo es importante que cumpla con una serie de requerimientos que harán más precisos los resultados que arroja, debe tener un buen rendimiento y funcionar adecuadamente aun cuando haya informacion faltante o errónea, debe ser claro al momento de realizar un diagnóstico y preciso al explicar las decisiones por las cuales lo deduce, el algoritmo debe ser capaz de reducir el número de pruebas necesarias para obtener un diagnóstico confiable.[13]

IV. Algoritmos

Existen diferentes algoritmos que son aplicados en el mundo del Data Mining y del Machine Learning, modelo de regresión logística (y otros modelos similares) es

extremadamente sensible a las respuestas periféricas y puntos extremos en el espacio de diseño.[18] Hay muchos factores a considerar a la hora de escoger el algoritmo más óptimo y preciso y esto va a depender exclusivamente del caso de estudio en el que se está trabajando.

Según los estudios realizados por Igor Kononenko en su paper "Machine learning for medical diagnosis: history, state of the art and perspective",[13] Naives Bayesian y el Semi-Naives Bayesian (que es una extensión del primero) son los algoritmos que mejores resultados devuelven en aplicaciones médicas.[13] En comparación con otros algoritmos muy usados como los árboles de decisión[9] o el k-nearest neighbours,[12] los Naives Bayesian pueden ser interpretados cómo la suma de la información que va obteniendo. Es decir, estos algoritmos van ganando información de cada uno de los atributos ya sea en favor o en contra de la clase dada.[19]

En el Cuadro I, se observa la comparación de diferentes algoritmos para el diagnóstico de cardiopatía isquémica realizados por Igor Konenko en su investigación.[13] En estos estudios se sugiere que los algoritmos Naives Bayes y Semi-Naives Bayes son lo que arrojan mejores resultados para la clasificación de nuevos casos usando como base para calcular la confiabilidad de casos nuevos todos los atributos.

Dicha investigación sirvió como base para la selección del algoritmo en nuestra implementación y cuyos resultados se muestran más adelante en el Cuadro II. Cabe aclarar que el presente trabajo abarca cualquier tipo de enfermedad coronaria, mientras que en el trabajo referenciado de Konenko[13] trata una única enfermedad coronaria en particular (cardiopatía isquémica). A continuación se describen brevemente los algoritmos mencionados en el Cuadro I.

IV-A. Naives Bayesian

Naives Bayes es uno de los algoritmos de aprendizaje más eficientes y efectivos dentro del Machine Learning y el Data Mining. Su rendimiento competitivo en la clasificación es sorprendente, porque la suposición de independencia condicional en la que se basa, rara vez es cierto en las aplicaciones del mundo real.[20]

Según estudio, este algoritmo fue comparado con otros 6 y en 5 de 8 casos para diagnósticos médicos, el Naives Bayesian obtuvo mejores resultados que todos los demás.

$$P(C|V..V_n) = P(C)\ IT\ \frac{P(C|V_i)}{P(C)} \qquad (1)$$

el mundo del Data Mining y del Machine Learning, y cada uno de ellos se adecua de diferentes maneras a los diferentes escenarios que se pueden presentar

en la vida real. Un ajuste de máxima probabilidad de un

Donde *P (X)* es la función de probabilidad que ocurra *X* y *P (X Y)* es la probabilidad condicional, i.e. la probabilidad que ocurra *X* sabiendo que también ocurre *Y* .

classifier	positive reliable (%)	cases erros (%)	negative reliable (%)	cases erros (%)
physicians	73	3	46	8
semi-naive Bayes (a)	79	5	46	3
Assistant-I (a)	79	5	49	8
neural network (a)	78	4	49	8
semi-naive Bayes (b)	90	7	81	11
Assistant-I (b)	87	8	77	6
neural network (b)	86	5	66	9
naive Bayes (c)	89	5	83	1
semi-naive Bayes (c)	91	6	79	2
Assistant-I (c)	77	18	55	18
Assistant-R (c)	81	5	77	2
k-NN (c)	64	12	80	12
neural network (c)	81	11	72	11

Cuadro I: Resultados de varios algoritmos clasificadores en Konenko.[13] El porcentaje de casos diagnoticados de forma confiable junto con los casos diagnosticados erróneamente es dado tanto para los casos positivos como para los negativos.

(a) Cálculo gradual de las probabilidades posteriores a la prueba.

(b) Usar todos los atributos a la vez para calcular las probabilidades posteriores a la prueba.

(c) Usar todos los atributos a la vez para evaluar la confiabilidad de la clasificación de casos nuevos únicos.

IV-B. Semi-Naives Bayesian

Es una extensión del clasificador Naive Bayesian que busca dependencias explícitas entre los diferentes valores de los distintos atributos.[20]

Si dicha dependencia entre dos valores es descubierta, entonces ellos no son considerados como condicionalmente independientes.

$$\frac{P(C|V_i)}{P(C)} \times \frac{P(C|V_j)}{P(C)} \qquad (2)$$

Esta parte (2) de la ecuación (1) es sustituida por (3):

$$\frac{P(C|V_i, V_j)}{P(C)} \qquad (3)$$

IV-C. Assistant-R

Assistant-R es una reimplementación del sistema de aprendizaje Assistant para la inducción descendente de árboles de decisión. La diferencia principal entre Assistant y su reimplementación Assistant-R es que la segunda usa ReliefF para la selección de atributos, la cuál es una medida heusística que puede estimar la calidad de los atributos, incluso si hay fuertes dependencias condicionales entre los mismos.[13]

I V-D. Assistant-I

Assistant-I es una variante de Assistant-R que en lugar de usar ReliefF, usa información ganada para el criterio de selección tal cómo lo hace el Assistant original.[13]

IV-E. k-NN *(k-nearest neighbor)*

El algoritmo k-nearest neighbor es un algoritmo supervisado basado en instancia. Puede usarse para clasificar nuevas muestras (valores discretos) o para predecir (regresión, valores continuos). Sirve para clasificar valores buscando los puntos de datos "más similares" aprendidos en la etapa de entrenamiento y haciendo conjeturas de nuevos puntos basados en esa clasificación.[21]

IV-F. Neural Network

Las redes neurales son un conjunto de algoritmos, modelados libremente según el cerebro humano que están diseñados para reconocer patrones. Interpretan lo datos sensoriales a través de una especie de percepción de máquina, etiquetado o agrupación de datos sin procesar. Los patrones que reconocen son numéricos, contenidos en vectores, a los que se deben traducir todos los datos del mundo real, ya sean imágenes, sonidos, texto o series de tiempo.[22]

Las redes neurales son realmente útiles para agrupar y clasificar.

V. Fuente de datos

Los datos utilizados fueron sacados de la Cleveland Database. Esta set de datos contiene 303 instancias con 75 atributos. De esos atributos, 14 son los que frecuentemente son tomados en cuenta para este tipo de investigaciones.[23]

Esta base de datos ha sido utilizada para otras investigaciones de Machine Learning.

Los 14 atributos principales de este set de datos son:

- Edad.
- Sexo.
- Tipo de dolor de pecho. Presión arterial en reposo. Colesterol sérico en mg/dl.
- Glucemia en ayunas (> 120 mg/dl) (1 = True; 0 = False).
- Resultados electrocardiográficos en reposo. Máximo ritmo cardíaco alcanzado.
- Angina inducida por ejercicio (1 = Sí; 0 = No). Depresión inducida por el ejercicio relativo al descanso.
- Pendiente del segmento de ejercicio pico.

Clasificador	Precisión Casos positivos (%)	Falsos positivos (%)	Precisión Casos negativos (%)	Falsos negativos (%)
Naive Bayes	88	12	85	15
	Precisión: 86,81 %		Error: 13,19 %	

Cuadro II: Resultados de la implementación del Naives Bayes para la muestra de 303 instancias. Se utilizaron 212 como parte del entrenamiento del algoritmo, y las 91 instancias restantes como parte del conjunto de prueba.

- Número de vasos principales coloreados por flourospía (0-3).
- Ritmo cardíaco.
- Diagnóstico de enfermedad del corazón.

Estos datos usados como set de entrenamiento para nuestro sistema nos proporcionaría una base confiable para la ejecución del mismo.

VI. Resultados y análisis

De las 303 instancias de nuestra fuente de datos, se utilizaron 212 instancias como parte del conjunto de entrenamiento y 91 instancias se utilizaron como

conjunto de prueba. Esto es un 70 % del conjunto inicial para instancias de entrenamiento del sistema y un 30 % para las instancias de prueba.

Usando esta distribución para la implementación inicial, obtuvimos un 86,81 % de precisión por parte de nuestro algoritmo (Cuadro II).

El conjunto de entrenamiento luego de ser proporcionado a nuestro algoritmo de aprendizaje (Naives Bayes), se le paso nuestro conjunto de prueba para poder realizar predicciones sobre los mismos. En base a este conjunto de prueba, se obtuvo el porcentaje de precisión mencionado anteriormente.

El 86,81 % de presición obtenido durante la implementación del sistema, está alineado con los resultados proporcionados por otros estudios en cuanto a la confiabilidad que el algoritmo puede proporcionar (Cuadro I). Hay que tomar en consideración que el set de datos utilizado no es lo suficientemente grande como para realizar un entrenamiento más exhaustivo del modelo, pero proporciona un escenario inicial que nos permite evaluar la confiabilidad del algoritmo seleccionado.

VII. Conclusión y trabajos futuros

En este trabajo, se analiza el uso de los algoritmos Naives Bayesian y Semi-Naives Bayesian para la clasificación de pacientes que sufren de enfermedades cardíacas y se compara su performance con diferentes algoritmos utilizados en el área de la medicina. Se muestra la gama de enfermedades que están incluidas dentro de las enfermedades cardíacas y cuales son las principales causas y síntomas para cada una de ellas. Los resultados sugieren que los algoritmos mencionados anteriormente son los más eficientes, sobre todo para trabajos médicos, y en que consiste cada uno. Se describe la fuente de datos que se usa para realizar una primera implementación y mostramos que, junto a resultados obtenidos en otras investigaciones, estos algoritmos brindan un nivel de confianza bastante alto, lo que nos permite establecer que los algoritmos de la familia Naive Bayes pueden ser ideales para implementaciones futuras en el campo de la medicina. Se va a continuar realizando estudios sobre estos algoritmos para buscar posibles alternativas en la implementación y buscando realizar estudios con set de datos más grandes para confirmar que con datos distintos y de mayor volumen, los resultados siguen siendo fiables. También se realizarán estudios sobre diferentes algoritmos para comparar los rendimientos de los mismos con pruebas más extendidas.

Referencias

[1] G. F. Tomaselli and D. P. Zipes, "What causes sudden death in heart failure?" *Circulation research*, vol. 95, no. 8, pp. 754–763, 2004.

[2] V. Autores, "Risk factors for heart diseases," url https://www.webmd.com/heart-disease/risk-factors-for-heart-disease.

[3] R. Vijayakrishnan, S. R. Steinhubl, K. Ng, J. Sun, R. J. Byrd,

[4] Z. Daar, B. A. Williams, C. Defilippi, S. Ebadollahi, and W. F. Stewart, "Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record," *Journal of cardiac failure*, vol. 20, no. 7, pp. 459–464, 2014.

[5] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.

[6] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson *et al.*, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature medicine*, vol. 7, no. 6, p. 673, 2001.

[7] A. Rajkumar and G. S. Reena, "Diagnosis of heart disease using datamining algorithm," *Global journal of computer science and technology*, vol. 10, no. 10, pp. 38–43, 2010.

[8] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[9] C. K. GN, "Machine learning types and algorithms," urlhttps://towardsdatascience.com/machine-learning-types-and-algorithms-d8b79545a6ec.

[10] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[11] H. Zhang, "The optimality of naive bayes," *AA*, vol. 1, no. 2, p. 3, 2004.

[12] G. I. Webb and M. J. Pazzani, "Adjusted probability naive bayesian induction," in *Australian Joint Conference on Artificial Intelligence*. Springer, 1998, pp. 285–295.

[13] Q. Kuang and L. Zhao, "A practical gpu based knn algorithm," in *Proceedings. The 2009 International Symposium on Computer Science and Computational Technology (ISCSCI 2009)*. Citeseer, 2009, p. 151.

[14] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in medicine*, vol. 23, no. 1, pp. 89–109, 2001.

[15] M. Clinic, "Enfermedad cardiaca," https://www.mayoclinic.org/es-es/diseases-conditions/heart-disease/symptoms-causes/syc-20353118.

[16] M. N. Today, "Heart disease: Types, causes and treatments," url https://www.medicalnewstoday.com/articles/237191.php.

[17] S. U. Amin, K. Agarwal, and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," in *2013 IEEE Conference on Information & Communication Technologies*. IEEE, 2013, pp. 1227–1231.

[18] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[19] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[20] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[21] M. J. Mizianty, L. A. Kurgan, and M. R. Ogiela, "Discretization as the enabling technique for the naive bayes and seminaive bayes-based classification," *The Knowledge Engineering Review*, vol. 25, no. 4, pp. 421–449, 2010.

[22] O. Harrison, "Machine learning basics with the k-nearest neighbors algorithm," https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm- 6a6e71d01761.

[23] C. Nicholson, "A beginner's guide to neural networks and deep learning," https://pathmind.com/wiki/neural-network.

[24] M. L. Repository, "Cleveland database - heart disease dataset," url http://archive.ics.uci.edu/ml/datasets/Heart+Disease.