4 -. Capítulo de Tesis: Diseño Metodológico

El presente capítulo tiene como finalidad el análisis de la documentación del trabajo realizado y la implementación del método utilizado. Para lograr un resultado óptimo primero se define que método se siguió. En la sección posterior se implementará los pasos comunes para los algoritmos, puesto que estos pasos serán frecuentes en los algoritmos que se utilizarán. Al finalizar, se presentarán nuevas secciones con los nombres de los algoritmos y los pasos no comunes del método dentro de estos.

\doublespacing

\section{Investigación}

\spacing{1.5}

Esta sección nos ayudará a comprender de qué forma científica estamos tratando de resolver el problema.\\

\doublespacing

\subsection{Enfoque o método de Investigación}

\spacing{1.5}

La metodología de este trabajo será de enfoque cuantitativa por los resultados que se quieren llegar a obtener, donde primará el análisis matemático de la base de datos y entre los algoritmos a comparar. \\

\doublespacing

\subsection{Tipo de Investigación}

\spacing{1.5}

El tipo de investigación será de carácter experimental, ya que medirá tendencias en los resultados arrojados por los algoritmos. \\

\doublespacing

\subsection{Población}

\spacing{1.5}

La población será todas las personas que salen registradas en la base de datos a trabajar. En este caso específico, son los pacientes post Ataque Cerebrovascular Isquémico del Hospital Herminda Martin de Chillán.\\

\doublespacing

\subsection{Muestra}

\spacing{1.5}

La muestra serán los pacientes post Ataque Cerebro Vascular Isquémico registrados en la base de datos y que cumplan con algunos criterios médicos para su análisis.\\

\doublespacing

\section{Enfoque de la Investigación}

\spacing{1.5}

La investigación cuenta con un enfoque cuantitativo por los resultados que se quieren llegar a obtener, donde primará el análisis matemático de la base de datos y entre los algoritmos a comparar. El tipo de investigación será de carácter experimental, ya que medirá tendencias en los resultados arrojados por los algoritmos.\\

\par La población será todas las personas que salen registradas en la base de datos a trabajar. En este caso específico, son los pacientes post Ataque Cerebrovascular Isquémico del Hospital Herminda Martin de Chillán. La muestra serán los pacientes post Ataque Cerebro Vascular Isquémico registrados en la base de datos y que cumplan con algunos criterios médicos para su análisis.\\

############ Desde aquí ################################

## 6.2 Método

El diseño metodológico dará una guía con los pasos a seguir a gran escala para la obtención de la finalización del proyecto. Se tendrá en cuenta el tipo, el enfoque, la población y la muestra para iniciar el trabajo. Siguiendo con la estrategia, utilizaremos el libro de \emph{Machine Leaning in Action} \cite{Harrington2012}, utilizando el procedimiento de la sección \emph{“Steps in developing a machine learning application”} que consta de 6 pasos para la implementación de una aplicación que utiliza técnicas de Machine Learning.

\begin{enumerate}

Colección de los datos de entrada: El primer paso para implementar una aplicación que trabaje utilizando técnicas de Machine Learning es coleccionar los datos que serán analizados, hay muchas formas en las que se pueden obtener grandes cantidades de datos a través de internet y de manera local, este paso se refiere a esas opciones y al esfuerzo aplicado para obtener dichos datos pudiendo ahorrar algo de tiempo y esfuerzo en adquirir datos públicos.

Preparación de datos de entrada: Una vez que ya se obtienen los datos, es necesario asegurarse que estén en el formato correcto para ser procesados por el algoritmo de Machine Learning seleccionado. El formato que usaremos en este estudio es la lista de Python. El beneficio de tener este formato estándar es que puede mezclar, combinar algoritmos y fuentes de datos. Es posible que se deba hacer algún formato específico del algoritmo aquí. Por lo que este paso involucra, si fuera necesario, formatear los datos para adaptarlos a la necesidad de cada algoritmo.

Analizar los datos de entrada: Este podría ser tan simple como mirar los datos que se ha analizado en un editor de texto para asegúrese de que los pasos 1 y 2 realmente funcionen y no tenga un montón de valores vacíos. También se podría observar los datos para ver si es posible reconocer algún patrón o si hay algo obvio, como también algunos puntos notoriamente diferentes del resto del conjunto. Si se trabaja con un sistema de producción o se sabe cómo se deben ver los datos o se confía en la fuente, este paso se puede omitir.

Entrenamiento del algoritmo: Aquí es donde tiene lugar el aprendizaje automático. Este paso y el siguiente es donde se encuentran los algoritmos "básicos", según el algoritmo. Alimentas el algoritmo con buenos datos, limpios de los primeros dos pasos y extrae conocimiento o información. Este conocimiento que a menudo almacena en un formato es fácilmente utilizable por una máquina para los siguientes dos pasos.

Testeo del algoritmo (Pruebe el algoritmo): En este paso la información aprendida por el algoritmo es testeada, es decir, se mide el nivel de acierto que tiene nuestro algoritmo. Utilizando los datos de entrenamiento, podremos establecer el grado de eficacia de nuestra implementación de la técnica seleccionada. Si los resultados no son los esperados es probable que se deba volver a etapas previas para identificar el error y tratar de solucionarlo, tal vez estos se encuentren en los datos de entrada o en el algoritmo en sí. Una vez realizado los cambios hará falta volver a pasar por todos los pasos anteriores una vez más.

Uso del algoritmo: Una vez se hayan ejecutado todos los pasos, no queda más que usar el algoritmo, esta etapa implica tener que volver a efectuar los pasos 1, 2, 3, 4 y 5. Sabemos que el proceso de programación podemos desarrollarlo óptimamente en los leguajes más usados para el Machine Learning que son Java, R o Python. Par este proyecto se escogió Python por su simpleza y múltiple documentación para el desarrollo de las técnicas de Machine Learning.\\

# Aplicación del método

Para empezar, debemos implementar los pasos sugeridos en la sección 4.1 en el mismo orden que se nos presenta. Los pasos 1, 2 y 3 son comunes para todos los algoritmos, por ende, solo se mostrará una vez en la presente investigación. Al finalizar esta sección se verán los pasos 4, 5 y 6 en cada Algoritmo presentado.

Colección de datos de entrada

Iniciamos obteniendo los datos desde la base de datos que facilitó el Hospital Herminda Martin de pacientes post ACV Isquémico. Esta base de datos fue trabajada por un grupo de profesionales que son expertos en la materia de ACV.

Se observa que la BDD posee muchas variables y pocos pacientes registrados en las tuplas. Esto hará que sea más difícil la predicción para los algoritmos, así que necesitamos un nuevo instrumento.

Diseño del instrumento

Desde el punto de vista científico, para que un estudio salga lo más certero posible necesitamos variables significativas para la investigación y con la menor perdida de datos posible. En la BDD todas las variables presentan importancia, algunas son imprescindibles para la investigación, otras con pocos datos completados o simplemente las variables sujetas a interpretación médica (humana). Debido a lo anterior se seleccionaron las variables por dos motivos, el primero fue porque eran las que estaban más completas en la BDD y el segundo porque se determinó que eran más significativas para la investigación por estudios realizados al ACV y dataset presentes en internet. A continuación, se mostrarán las variables escogidas y una pequeña descripción de ellas.

La Hipertensión Arterial y la Diabetes son factores de riesgo altos en cualquier enfermedad no trasmisible, por esto son de las primeras seleccionadas, que además contaremos con las escalas de Glasgow y Nihss que son escalas internacionales para la evaluación del ACV Isquémico.

##################################################################

# Preparación de los datos de entrada

El segundo paso descrito en la metodología corresponde a la parte más larga de la elaboración para la preparación de la utilización de un modelo de Machine Learning, debido a que este paso todos los datos serán usados como entrada del algoritmo. Aquí se llevará a cabo una identificación y corrección de los puntos de datos faltantes/anomalías según sea necesario.

Eliminación las filas de los pacientes que se expulsaron de la BDD

Como se indicó anteriormente, se dispone de 75 tuplas con 85 columnas, con un total de 6375 elementos, que se planean disminuir por indicación del médico que facilitó la base de datos. La indicación fue que había pacientes que fueron retirados del programa y estaban marcados con un “out” en la variable de “diag Elopez”.

Variables significativas para la investigación

Ahora asignamos las variables significativas, para esto se extrae la información del marco de trabajo, identificando las variables categóricas para un arreglo completamente nuevo y asi empezar a trabajar sobre el nuevo archivo.

Descripción general de los datos

En la descripción de los datos, se muestran parámetros, perdida de datos (Missing Data), forma y su descripción estadística.

Missing data

Los datos que faltan ocurren cuando no se almacena ningún valor en la variable de observación. Aunque el Missing data es una ocurrencia muy común, estos pueden tener una presión significativa en los resultados de la aplicación del instrumento. Para corregir este problema, existen variadas técnicas estadísticas, siendo una de ellas la mediana y el redondeo como lo muestra Will Kushner en algunas variables que utilizaremos para llenar las celdas faltantes. Tomaremos la información que mostramos anteriormente.

Preprocesamiento de los datos y clasificación

La preparación de los datos para el modelo de aprendizaje automático es lo que nos llevará a decidir qué modelo podremos ocupar, ya que trabajar para modelos de clasificación no es lo mismo que para los de regresión. Los pasos exactos para la preparación de los datos dependerán del modelo utilizado y de los datos recopilados, pero requerirá cierta cantidad de manipulación de datos para cualquier aplicación de aprendizaje automático.

Análisis de variable Objetivo

La NIHSS es la escala más utilizada para evaluar la función neurológica básica durante la fase aguda del ictus isquémico, tanto en su inicio como en su evolución. Consta de 11 ítem que te permiten explorar rápidamente (Cién escalas de interés en Neurología. Prous Science,2001). Según las puntuaciones obtenidas, podemos dividir la gravedad neurológica en varios grupos: 0: sin déficit; 1: déficit mínimo; 2-5: leve; 6-15: moderado; 15-20: déficit importante; >20: Gravemente ( Montaner 2006). La puntuación global inicial tiene buen valor de pronóstico (Heinemman 1997) (Adams 1999), dado que NIHSS ≤ 6 corresponde a una excelente recuperación neurológica y que cada aumento de punto empeora la evolución. Para los pacientes con fibrilación auricular, se ha considerado que un NIHSS ≥16 tiene un pronóstico muy precario (Frankel 2000). Otra ventaja de esta escala es que puede predecir la respuesta a la terapia trombolítica y estimar la gravedad en función de la puntuación.

La variable más importante que nos interesa tener en la predicción, es la escala Nihss de alta, esto para poder tener un aseguramiento de la predicción del estado de salud del total de pacientes a través del camino de salida del hospital, es decir, conoceremos el estado del paciente al salir de alta y según las demás variables existe una tendencia para la predicción en la escala del alta.

#### NIHSS INICO ACV

Esta variable la transformaremos a la categoría que nos ofrece la literatura con sus 6 estados.

La realización de la escala fue acorde a los mismos parámetros que la variable objetivo, solo que en este caso dejamos de lado la variable binaria.

Se observa que el valor mínimo que llego un paciente fue de 0 encontrándose Sin Déficit neurológico y el máximo con 21 puntos con un Déficit importante, además se aprecia gracias al grafico que los pronósticos moderado y leve son los que más abundan en la muestra, al mismo se demuestra que esos pacientes al salir de alta en su mayoría tienen un buen pronóstico según la escala de NIHSS.

#### GLASGOW AL INICO ACV

La escala de Glasgow se divide en tres grupos puntuables de manera independiente que evalúan la apertura de ojos sobre 4 puntos, la respuesta verbal sobre 5 y la motora sobre 6, siendo la puntuación máxima y normal 15 y la mínima 3 (The status of the Glasgow Coma Scale). Se considera traumatismo craneoencefálico leve al que presenta un Glasgow de 15 a 13 puntos, moderado de 12 a 9 y grave menor o igual a 8 (Cién escalas de interés en Neurología. Prous Science,2001). A continuación, procesaremos la escala.

Se observa que el valor mínimo que llego un paciente fue de 0 encontrándose grave neurológico y el máximo con 15 puntos encontrándose leve, además se aprecia gracias al grafico que el pronostico Leve es el que abunda en esta variable, al mismo se demuestra que esos pacientes al salir de alta en su mayoría tienen un buen pronóstico según la escala de NIHSS.

Análisis de otras variables de médicas

Las siguientes clasificaciones están sacadas del (<https://medlineplus.gov/spanish/ency/esp_presentations/100151_3.htm>) y Geo Salud, que son páginas que poseen escalas para la clasificación de las variables que veremos.

#### Conteo de Glóbulos Blancos

Los glóbulos blancos son parte del sistema inmunológico del cuerpo y ayudan a combatir infecciones y otras enfermedades. Los registraremos en 3 categorías.

Se observa en el grafico que la mayoría de los pacientes están en la categoría normal y ellos tienen en más del 50 \% un buen pronóstico.

#### INR

El INR sirve para ver la coagulación de la sangre, siendo que esta examen ayuda a saber qué tipo de medicamentos podemos administrar o averiguar la causa de los coágulos sanguíneos anormales.

Se observa en el grafico que la mayoría de los pacientes están en la categoría de Riesgo, aunque un grupo importante tiene un pronostico favorable. En la categoría Normal no se sabe si el paciente obtuvo un buen o mal pronóstico.

#### Triglicéridos

Los triglicéridos son el principal tipo de grasa que se transporta en la sangre para obtener energía o se almacena en las células del cuerpo según las necesidades energéticas entre comidas.

El grafico muestra que la mayoría de los pacientes están en la categoría Normal, aunque un grupo importante tiene un pronóstico favorable. En la categoría Alta un gran porcentaje tiene un mal pronostico con el alta del ACV.

#### Colesterol Total

El colesterol es una grasa natural que se encuentra en todas las células del cuerpo y es necesaria para que el cuerpo funcione correctamente. El colesterol se produce en el hígado en su mayoría, aunque también se puede obtener de algunos alimentos.

El grafico muestra que la mayoría de los pacientes están en la categoría Limite alto, siendo que un grupo importante tiene un pronóstico favorable.

#### Glucosa

El azúcar en la sangre, es llamada "glucosa", siendo el azúcar principal en la sangre. Proviene de los alimentos que consume y es la principal fuente de energía. Su sangre transporta glucosa a cada célula de su cuerpo para obtener energía.

El grafico muestra que la mayoría de los pacientes están en la categoría Normal, aunque un grupo importante tiene un pronóstico poco favorable en el alta. En la categoría Alta un gran porcentaje importante tiene un buen pronóstico con el alta del ACV.

#### Edad

La edad juega un rol importante para determinar el estado que se puede encontrar el cuerpo para responder a una enfermedad o tratamiento, es por eso que una de las variables que si o si debe estar presente.

El grafico muestra que la mayoría de los pacientes que son ancianos pasaron por el ACV, aunque un grupo mayoritario de esa categoría tiene un pronóstico favorable en el alta. En la categoría Adulto es la segunda más grande y si respuesta hacia el alta es muy pareja en los dos pronósticos.

#### Diabetes

La diabetes es una enfermedad crónica que afecta la forma en que el cuerpo convierte los alimentos en energía. El cuerpo descompone la mayor parte de los alimentos que ingiere en azúcar (también llamada glucosa) y la libera en la sangre. Puede existir exceso de azúcar en la sangre o ausencia de azúcar en la sangre.

En este caso, solo se tomo si el paciente poseía diabetes y los datos desconocidos se tomó como otra variable.

El grafico muestra que la mayoría de los pacientes que tienen diabetes tienen un buen pronóstico al alta. El estado Desconocido posee un pronóstico poco favorable al alta.

#### Hipertensión

La hipertensión es sinónimo de enfermedades cardiacas, siendo que es la fuerza de la presión arterial que trasporta los nutrientes con la sangre.

El grafico muestra que los pacientes que poseen hipertensión son muchos y su pronóstico es similar a que estado pueden tomar.

Binary Encoding (Codificación binaria)

Lo que hace este proceso es la transformación de variables de una columna a binario si es posible, esto solo aplica a las viables que tiene dos estados, si posee más estados no podrá aplicar.

Label Encoding (Codificación de etiquetas)

Estas son variables ordinales, la codificación consiste en la traducción de ese orden en una enumeración, donde cada categoría es asociada con un número natural respetando el orden implícito existente entre categorías (J. T. Hancock and T. M. Khoshgoftaar, “Survey on categorical data for neural networks").

## One-Hot Encoding:

Para las variables categóricas donde no existe tal relación ordinal, la codificación de enteros no es suficiente. De hecho, usar esta codificación y permitir que el modelo asuma un ordenamiento natural entre categorías puede dar como resultado un desempeño deficiente o resultados inesperados (predicciones a mitad de camino entre categorías) ( “Sparse matrix multiplication package (smmp),” ). En este caso, se puede aplicar una codificación one-hot a la representación de enteros. Aquí es donde se elimina la variable codificada entera y se agrega una nueva variable binaria para cada valor entero único. En el ejemplo de la variable " color ", hay 3 categorías y, por lo tanto, se necesitan 3 variables binarias. Se coloca un valor "1" en la variable binaria para el color y valores "0" para los otros colores.

######################################################

**4.5. Análisis de datos de entrada**

Ahora pasaremos al punto 3 de la metodología, veremos gráficos y datos que

interactuaran entre si. Nuestro interés por esta sección radica por la observación

de alguna tendencia de los datos. Crearemos dos dataset auxiliar, para observar los

datos que estaban y los que incorporamos en la seccion anterior.

Se observa que ambos dataset tienen un numero de cantidades de variables similares y que la última posee datos más cercanos a lo binario.

Análisis de densidad y estimación por variable

La densidad de los datos nos da la probabilidad de que una variable tomé un valor determinado por su densidad. La importancia de la densidad radica en comprender en que valores los datos están siendo más remanentes.

Como los datos son variados numéricamente, en el grafico no se alcanza a reflejar cada variable. Si fuera necesario deberíamos analizar en este caso cada variable por separado, siendo que este no es el caso para la mayoría de las variables. A continuación, se mostrará el grafico de densidad de los exámenes.

Las tendencias de los gráficos están orientadas al grupo etario entre los 60 y 80 años estos que poseen una densidad en los datos alta, es decir que los exámenes se encuentran semejantes a los datos.

Se puede interpretar el grafico de la siguiente forma: En la variable “NIHSS\_INICIO\_cat\_Sin Déficit” la densidad de los datos es alta en el valor 0, entendiendo que esta poseía solo dos estados, 0 para confirmar que no está presente el valor y 1 para confirmar que está presente. Si lo traducimos a palabras normales, los pacientes sin déficit al principio medidos por la escala son muy pocos, la mayoría presenta déficit. El análisis individual no es necesario en este caso, ya que los valores están bien detallados a la vista, aunque estos son los valores importantes que dejo el procesado de datos de entrada.

Análisis por conteo de variables categóricas

Para demostrar la cantidad de datos presente en algunas de las variables originales que son más agrupables se representará con escalas de contador por categoría.

Las tasas de conteo de los pacientes indican una tendencia en algunos valores específicos, donde a simple vista según las escalas los pacientes se encuentran en buen estado, pero la gran mayoría contiene enfermedades categóricas no transmisibles.

Distribución de Diabetes e Hipertensión

La distribución de la Diabetes e Hipertensión es inversa en la escala de alta de los pacientes. El grafico muestra que en su mayoría los pacientes que poseen diabetes tienen un pronostico al alta más favorable, en cambio los que poseen hipertensión tienen un pronóstico menos favorable. Sin embargo, la distribución de pacientes es muy pronunciada hacía que los pacientes poseen las dos enfermedades en conjunto.

Mapa de calor de variables

Los mapas de calor pueden ayudar a visualizar el impacto de muchas categorías en los valores y son una excelente opción para ver detalles cuando se trabaja con conjuntos de datos más grandes. Los mapas de calor utilizan una variedad de colores fríos y cálidos para ayudarnos a comprender qué elementos de datos generan más interés (áreas calientes) y qué elementos se ignoran (áreas frías).

El primer dataset muestra un alto interés en casi todas las variables tendiendo al rojizo y algunos sectores fríos, en cambio el segundo dataset presenta variables más estables con estados más cercanos a la neutralidad.

# Naïve Bayes - Entrenamiento del algoritmo

Este algoritmo es utilizado para predecir grandes volúmenes de datos. En este caso no se cuenta con un gran volumen de datos. El clasificador Naïve-Bayes aprende de los datos de entrenamiento y luego predice la clase de la instancia de prueba con la mayor probabilidad posterior. También es útil para datos dimensionales altos ya que la probabilidad de cada atributo se estima independientemente (Maquinas de soporte vectorial y anive bayes y algoritmos geneticops).

En el entrenamiento del algoritmo el programa de Machine Learning adquiere la información que trabajamos en los métodos anteriores. Es aquí donde se obtendrá el conocimiento para futuras decisiones, es importante asegurarse que las decisiones que sean tomadas posteriormente al proceso de entrenamiento se añadan a la base de conocimiento del algoritmo para futuras ejecuciones de este.

La BDD trabajada actualmente cuenta con: La BDD trabajada actualmente cuenta con:

Variable categórica

En el paso de la preparación de los datos de entrada, propusimos la variable “NIHSS alta ACV” que podía poseer 42 valores diferentes, la cual se clasificó y se transformó en “NIHSS\_alta\_cat” que contenía 6 categorías las que fueron reducidas a 1 variable con dos estados. Asi el paciente tendrá un buen pronóstico o no con el nombre de la variable “NIHSS\_alta\_ESTABLE\_O\_GRAVE”.   
Los estudios de Machine Learning no sugieren tener variables binarias para nuestro estudio.

Las variables dependientes representan el rendimiento o conclusión que se está estudiando. Las variables independientes, además conocidas en una relación estadística como regresores, representan insumos o causas, donde se encuentran las razones potenciales de alteración.

Creación del modelo y entrenamiento

Para la creación del modelo se utilizará el modelo en la forma más estándar posible, siendo que los modelos, antes del entrenamiento, pueden recibir ajustes para manejar los datos de entrada, de una forma u otra. Para que sea lo más parejo posible entre modelos se dejará de forma estándar.

Predicciones sobre los datos de prueba y métricas de rendimiento

Para llevar una forma más ordenada, es necesario crear las variables de predicciones, para asi sacar las métricas de rendimiento más fácilmente. Las métricas de rendimiento nos ofrecerán información de cómo se comportó el algoritmo durante el entrenamiento, dando a conocer valores importantes como lo son la precisión, exhaustividad, valor-F.

La precisión de los datos de entrenamiento en el modelo tiene un valor excelente de 100% de predicción, la exhaustividad informa la cantidad de datos capaz de identificar y en este caso es de un 100% de los datos y finalmente el F1 combina los valores de precisión y exhaustividad obteniéndose un 100% igual. Todos los valores mencionados aplican para los estados de la variable predictora.

Matriz de Confusión

En el campo de la inteligencia artificial y en especial en el problema de la clasificación estadística, una matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado.

En la matriz de confusión (1, 1) podemos observar el resultado en el que el modelo predice correctamente la clase positiva y en el (2, 2) el resultado donde el modelo predice correctamente la clase negativa. Los demás elementos de la matriz contienen valor nulo o 0, estos son los errores de la predicción.

######################################################################

# Clasificador Naive Bayes - Testeo del algoritmo

El testing tiene la finalidad de llevar a cabo la prueba si el modelo funciona correctamente, identificando riesgos o erros que se produjeron en los datos. No se realizará ajustes posteriores al testing para poder comparar los algoritmos en la sección de resultados.

Predicciones sobre los datos del testing y métricas de rendimiento

Ahora es momento de evaluar los datos ya entrenados con el testing. Las métricas de rendimiento nos ofrecerán información de cómo se comportó el algoritmo durante el entrenamiento, dando a conocer valores importantes como lo son la precisión, exhaustividad, valor-F.

La precisión de los datos del testing en el modelo tiene un valor de 59% de predicción para el estado 0 y un 60% de predicción para el estado 1. La exhaustividad en el estado 0 alcanza el 80% de los datos y en el estado 1 alcanza solo el 35%. Por otra parte, el F1 combina los valores de precisión y exhaustividad obteniéndose un 68% en el estado 0 y un 44% en el estado 1.

Lo que se busca es la precisión del modelo, por consecuencia, el Algoritmo de Machine Learning Logistic Regression tiene una precisión del 59,4% de predicción.  
###Matriz de Confusión

Evaluaremos la matriz de confusión que se elaboró con los datos del testing.

En la matriz de confusión (1, 1) podemos observar el resultado en el que el modelo predice correctamente la clase positiva con un alto valor y en el (2, 2) el resultado donde el modelo predice correctamente la clase negativa también con un alto valor, acercándose a la misma de la clase positiva. Los demás elementos de la matriz contienen valor pequeño, estos son los errores de la predicción.

Las afirmaciones anteriores sugieren que la las predicciones son altas, pero también existen errores en la predicción.

# Clasificador Naive Bayes - Uso del algoritmo

El último paso de la metodología es el uso del algoritmo, nosotros lo utilizaremos para desarrollar probabilidades en los predictores. No todos los modelos poseen los mismos métodos ni atributos, por ende, se tratará de realizar comparaciones con métodos similares entre sí.

Importancia de los predictores

Por experiencia previa y contemplando los gráficos producidos en el paso 3, sabemos que algunas características no son útiles para nuestro problema de predicción. Reducir la cantidad de funciones será la mejor alternativa, lo que acotará el tiempo de ejecución, con suerte sin comprometer significativamente el rendimiento, asi podemos examinar la importancia de las funciones de nuestro modelo. La importancia de cada predictor en el modelo se calcula como la reducción total (normalizada) en el criterio de división. Si un predictor no ha sido seleccionado en ninguna división, no se ha incluido en el modelo y por lo tanto su importancia es 0.

Este método acepta un solo argumento que corresponde a los datos sobre los cuales se calculan las probabilidades y devuelve una matriz de listas que contienen las probabilidades de clase para los puntos de datos de entrada.

En este caso particular podemos observar que los estados de la variable predictora tienen un valor de porcentaje predictor, por ejemplo, la tupla 0 posee un 99% y fracción de precisión para el estado 0 y un 0,11% y fracción para el estado 1.  
En síntesis se observa que cada uno de las variables reaccionan y toma valores porcentuales a que estado del predictor pertenece.

Se observa un valor binario de 0 o 1, donde se muestra cada variable desarrollada en el modelo puede tomar dicho valor. El valor 0 demuestra que la tupla no logra predecir el estado 0 de la variable predictora, y por el contrario, el estado 1 es que logra la predicción del estado en esa tupla.

##############################

Las variables que presentan gran importancia al momento de la predicción son “NIHSS\_alta\_cat”, “NIHSS alta ACV”, “CONTEO G.B.\_cat” y “GLUCOSA\_cat” con más del 98%, siendo que el mayor predictor esta con un 41,9% que es “NIHSS\_alta\_cat. Se ve reflejado en el grafico la gran importancia que representan las variables mencionadas para el modelo.

En el gráfico de importancia acumulada la curva se dispara con la “GLASGOW AL INICO ACV” y se ratifica con el resultado del 95% de importancia medido anteriormente con el valor cercano del 95% de importancia acumulada de los datos.

El cálculo de importancia muestra que existen muchas variables con importancia al momento de la predicción, siendo la que destaca “TRIGLICERIDOS” con un 14,9% de importancia. Viendo el grafico, se observa que aún hay variables que no son significantes.

Arboles:::::

Al observar lo tabla de valoración y los gráficos, se hace notar que las variables más importantes para la predicción en este modelo fueron los Triglicéridos y la escala NIHSS de alta.

El grafico muestra el alza de la importancia acumulada y su peak.

## 6.2 Procedimientos de la Investigación

**6.2.1-. Dominio del problema**

Dentro del análisis de los algoritmos, debemos dominar el problema revisando la bibliografía relacionada con el área que realizaremos la investigación, así como son los enfoques del Machine Learning (informática) y los ACV (salud), en definitiva, estas dos áreas se juntan para dar lugar a los modelos y tecnologías actuales que se implementan en ambas áreas o por separado, en el ámbito nacional o internacional (Objetivo Específico 1).

**6.2.2-. Creación del modelo**

En caso de no existir modelos de Machine Learning con ACV, se investigarían modelos parecidos para contrastar lo introducido en un nuevo modelo a la ciencia. A fin de todo esto habrá que elegir un mecanismo adecuado de Machine Learning (Objetivo Específico 2). En este punto también constará con un modelo como arquitectura, definiendo un marco claro para la interacción con el código fuente del software (Objetivo Específico 3).

**6.2.3-. Implementación: Construcción y eficiencia del Modelo**

Una vez creado el modelo, se baja hasta el cuadro de Implementar y realizaremos lo que son la recolección de datos. Los datos serán de tipo etiquetado, como datos algoritmos, datos estadísticos, entre otros, que nos ayuden al entrenamiento de la red neuronal, en base a la arquitectura planteada. Por consecuencia, la estrategia para la creación de la red neuronal, será aprendida desde el marco teórico para la creación del modelo, siendo que los datos deben ser procesados y ajustados al modelo paralelamente. esos datos. A continuación, una vez finalizadas las tareas antes mencionadas, vendrá la tarea de Visión Data, que consiste en analizar un problema de perdida de información cuando la base de datos no está completa y para eso debemos estandarizar los datos y aplicar una estadística de normalización.

**6.2.4-. Experimentación del Modelo**

Para tener un buen modelo debemos realizar, en primera instancia, la tarea de prueba para ver si funciona o no, con casos clínicos que fueron seguidos durante un transcurso prolongado de tiempo, luego tendríamos la tarea de validación, que sería si logra completar con lo que se solicita para ser utilizado el modelo por otros profesionales.

**6.2.5-. Resultados de la Investigación**

Los Resultados son la parte más importante de la investigación, en donde la tarea será reflejar lo que obtuvimos y deseábamos, o simplemente tener unas métricas desde otro modelo y compararlas con nuestros resultados. Se destaca que el estado del Arte del capítulo 6, nos ayudar a tener una comparativa importante con los resultados de modelos actuales y con los resultados que entrega el modelo creado.

Después de todo lo anterior, las conclusiones serán la interpretación de los resultados y la investigación.

Los estudios experimentales se realizarían en el Hospital Herminda Martin, con pacientes que estén pasando por un ACV.

DESCRIPCIÓN DEL AMBIENTE DE SOFTWARE (SI CORRESPONDE)

## 6.3 Técnicas para la Recolección de Datos

## 6.3.1 Descripción de los Instrumentos

## 6.3.1.1 Representación de los datos

Los datos estarán dispuestos en una base de datos estática tipo hoja de cálculo, siendo que pueden ser transformado a ficheros tipo CVS que son más cercanos a la librería pandas, para poder ser usados por modelos de predicción en Python.

## 6.3.1.2 Elección de variables significativas

-Estimación del ritmo.

- Frecuencia Respiratoria.

## 6.3.1.3 Datos y normalización

Con la base de datos en nuestro poder, ahora es tiempo de procesar los datos en búsqueda de ausencia de datos (valores nulos).

* Normalización por Media y Desviación Estándar
* Normalización por Máximos y Mínimos

## 6.3.2 Validez y Confiabilidad de los Instrumentos

## 6.4 Técnica para el Procesamiento y Análisis de Datos

Entrenamiento

## 6.5 Resguardos Éticos

SI ESTE ANTEPROYECTO HA CONTADO CON LA COLABORACIÓN DE PERSONAS O INSTITUCIONES PARA LAS CUALES, FINALMENTE SERÁ DESARROLLADO EL PROYECTO, POR FAVOR IDENTIFICAR:

# Capítulo 7 Resultados

## 7.1 Predicción del tipo de Accidente Cardiovascular

Aquí se explica que cual de los algoritmos es más efectivo para predecir que es Isquémico.

## 7.2 Predicción de pronóstico

Debería haber datos de monitorización y de acuerdo a eso ver si el paciente vivirá o morirá. Tiene que ver más con los signos.

## 7.3 Predicción de aumento o disminución del total de variables

Aquí se hará un análisis del paciente para saber cuanto tiempo se tardará en recuperarse lo antes posible de todos los síntomas.

## 7.4 Predicción de reincidencia

Aquí se evaluará si el paciente volverá a ser reincidente. Depende de la base de datos a analizar, se podría poner en contraste los síntomas de antes vs los de la segunda vez.

# Capitulo 8 Discusión, Conclusiones y Recomendaciones

FIN :)

Relleno

\par El ACV posee clasificaciones y que se actualizan constantemente, ya que es una enfermedad grave que conlleva alteraciones significativas en la vida del paciente. Además se puede enfatizar que existe un confiable sistema de clasificación, debido a que este comprende un 75,2\% de las muertes y el 81\% de los años de vida ajustados por discapacidad que genera esta enfermedad en los países en desarrollo, asimismo existe un 87\% con el ACV Isquémico \cite{Radu2017}, ) son una clasificación confiable y precisa para la enfermedad a través del TOAST, aunque la clasificación fenotípica ha avanzado a pasos agigantados, aún se necesita más investigación a nivel mundial para lograr un lenguaje más estándar en la investigación Fenotípica.\\