**UNIVERSIDAD DE SANTIAGO DE CHILE**
**FACULTAD DE INGENIERÍA**
**DEPARTAMENTO DE INGENIERÍA INFORMÁTICA**

# INFORME LABORATORIO 1:

# MAGIC GAMMA TELESCOPE DATA SET

**DIEGO POLANCO BERRIOS, JUAN ROA CARVAJAL, DANY RUBIANO JIMENEZ**

Profesores:

- Mónica Villanueva Ilufi
- Felipe Bello Robles

Santiago - Chile

28 de marzo de 2016

# TABLA DE CONTENIDOS

# ÍNDICE DE FIGURAS

# ÍNDICE DE CUADROS

# CAPÍTULO 1. INTRODUCCIÓN

## 1.1 MOTIVACIÓN Y ANTECEDENTES

MAGIC proviene de Major Atmospheric Gamma-ray Imaging Cherenkov Telescope, es decir, ”Telescopio Cherenkov de rayos gamma por emisión de radiación en la atmosfera”. MAGIC es un telescopio con el cual es posible captar los destellos de luz producidos en la atmósfera terrestre proveniente de rayos cósmicos (en la galaxia). La principal función que tiene este telescopio es captar rayos gamma de muy alta energía (rayos mayores a 10 GeV). Los rayos gamma de más alta energía son capaces de producir extensas ”lluvias” con una forma cilíndrica en la atmósfera. La partículas que son producidas en estas lluvias emiten destellos de luz Cherenkov en un tiempo ínfimo (pocos billonésimos de segundos). Esos destellos pueden ser de cierta forma ”fotografiados” con los telescopios Cherenkov. El procesamiento de las imágenes mediante distintos tipos de métodos puede suprimir el fondo compuesto por rayos cósmicos hadrónicos. Para su funcionamiento el telescopio MAGIC cuenta con una estructura de 17 metros de diámetro construida con fibra de carbono tubular que cuenta con la propiedad de ser ligera además de rígida. Con esto el telescopio es capaz de reposiciones de forma rápida luego de una alarma proveniente del sistema de satélites que monitorizan las explosiones de rayos gamma. El sistema de espejos esta formado por 1000 espejos de un tamaño de 50cm x 50cm. A su vez, cada espejo consiste en una lamina de aluminio pulido con diamante, además de un sistema de calefacción para protegerlo de la condensación y el hielo, y una base estructural con una forma similar a un panel de abejas para entregar rígidas al telescopio. MAGIC es una herramienta potencial para realizar descubrimientos en Astrofísica, Cosmología y Física de Partículas.

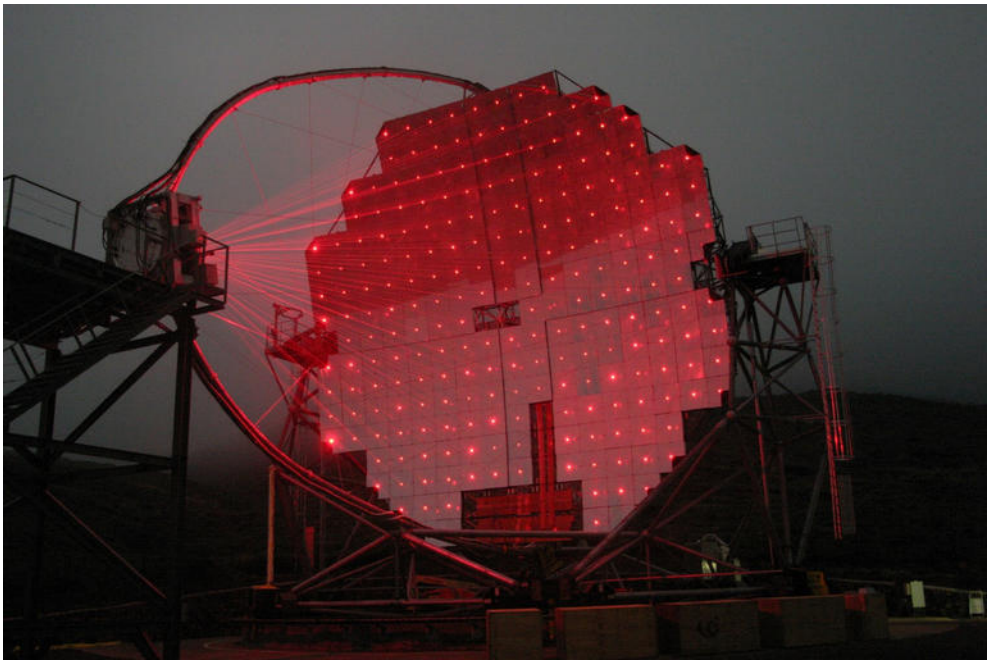

*Figura 1-1: Telescopio MAGIC*

Las técnicas multivariantes son un conjunto de métodos del tipo estadísticos que tienen como finalidad analizar de forma simultanea grupos de datos del tipo multivariante, es decir, varias variables a medir para cada caso. Estas técnicas permiten un mayor entendimiento del fenómeno estudiado, obteniendo de esta forma información que en métodos univariantes y bivariantes no son capaces de conseguir.

## 1.2 OBJETIVOS

El objetivo de este documento es estudiar los datos otorgados en un dataset provenientes del telescopio MAGIC. Se describen sus atributos, clases y sus valores. Con el respectivo estudio se adquiere una idea clara y precisa del trabajo que se está desarrollando. En especifico en este documento se desea dilucidar cuál de las técnicas que son descritas posteriormente, es óptima para el procesamiento de las imágenes del telescopio MAGIC, es decir, que método optimiza la diferenciación de rayos cósmicos hadrónicos (fondo) de los rayos gamma de interés (señales).

## 1.3 ORGANIZACIÓN DEL DOCUMENTO

El presente documento distribuye su contenido de la siguiente forma: primero se encuentra una descripción del problema, donde se presenta la información considerada relevante para la comprensión de la problemática. A continuación se realiza una descripción de la base de datos con la que el grupo trabaja, la descripción de las clases y variables respectivas de la base de datos y también lo denominado ”Estado del Arte”, que corresponde a los métodos empleados en el análisis de los datos. Con esta descripción del problema, del contexto que lo envuelve y de los datos obtenidos y presentados en la base de datos, se realiza el análisis de interés el cual es presentado a en la conclusión del presente informe.

# CAPÍTULO 2. GLOSARIO

**Multivariado:** Se refiere a un tipo de método estadístico en los cuales se pueden analizar simultáneamente la relación existente entre variables correlacionadas.

**Redes neuronales (ANN):** Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas que colaboran entre sí para producir un estímulo de salida. En inteligencia artificial es frecuente referirse a ellas como redes de neuronas o redes neuronales.

**Métodos Kernel:** Son una familia de algoritmos relativamente nueva en el mundo del análisis inteligente de datos y reconocimiento de patrones. Combinan la simplicidad y eficiencia de algoritmos como el perceptron y ridge regression con la flexibilidad de sistemas no-lineales y el rigor de los métodos de regularización

**Vecino más cercano:** Es un método para clasificar casos basándose en su parecido a otros casos. En el aprendizaje automático, se desarrolló como una forma de reconocer patrones de datos sin la necesidad de una coincidencia exacta con patrones o casos almacenados. Los casos parecidos están próximos y los que no lo son están alejados entre sí. Por lo tanto, la distancia entre dos casos es una medida de disimilaridad.

**Árboles de regresión:** Son una técnica de análisis discriminante no paramétrica que permite predecir la asignación de muestras a grupos predefinidos en función de una serie de variables predictoras. Es decir, que teniendo una variable respuesta categórica, los árboles de regresión permiten crear una serie de reglas basadas en variables predictoras que a suvez van a permitir asignar una nueva observación a un grupo u a otro.

**Análisis del discriminante:** Es una técnica estadística multivariante cuya finalidad es describir (si existen) las diferencias significativas entre g grupos de objetos (g ¿1) sobre los que se observan p variables (variables discriminantes). Más concretamente, se comparan y describen las medias de las p variables clasificadoras a través de los g grupos.

**Simulación Monte Carlo:** Es una técnica matemática computarizada que permite tener en cuenta el riesgo en análisis cuantitativos y tomas de decisiones. La simulación Monte Carlo ofrece a la persona responsable de tomar las decisiones una serie de posibles resultados, así como la probabilidad de que se produzcan según las medidas tomadas. Muestra las posibilidades extremas —los resultados de tomar la medida más arriesgada y la más conservadora— así como todas las posibles consecuencias de las decisiones intermedias.

**Boostrap:** El bootstrap es un tipo de técnica de remuestreo de datos que permite resolver problemas relacionados con la estimación de intervalos de confianza o la prueba de significación estadística.

**Evento:** Suceso o acontecimiento captado por el Telescopio Cherenkov que puede ser gamma (señal) o hadrón (trasfondo).

**Hadrón:** Es una partícula subatómica formada por quarks que permanecen unidos debido a la interacción nuclear fuerte entre ellos. Antes de la postulación del modelo de quarks se definía a los hadrónes como aquellas partículas que eran sensibles a la interacción fuerte.

**Señal Gamma:** Tipo de radiación electromagnética, producida por elementos radioactivos o partículas subatómicas.

**GeV o Gigaelectronvoltio:** Es una unidad de medida de energía, que representa la energía que experimenta un electrón cuando se mueve de un punto a otro, que equivale a $1,602176565 * 10^{-}10J$.

**Telescopio Cherenkov:** Un telescopio Cherenkov es un detector de rayos gamma de muy alta energía en el rango de 25 GeV a 50 TeV desde la superficie terrestre.

# CAPÍTULO 3. DESCRIPCIÓN DEL PROBLEMA

En el capítulo presentado a continuación se realiza la descripción detallada del problema que se aborda y, posteriormente, se analiza gracias a la información entregada en la base de datos correspondiente.

El problema, en este caso, es lograr la correcta identificación de los tipos de rayos (rayos gamma o hadrónicos) en una imagen captada mediante el telescopio MAGIC. Esto es especialmente útil en el campo de la Astrofísica, Cosmología y Física de Partículas, donde la correcta identificación puede ser determinante en el estudio del origen de los rayos cósmicos (hadrónicos),el origen de las explosiones de los rayos gamma, naturaleza de la materia oscura, pulsares y en el estudio de los agujeros negros súper masivos.

Para el proceso de identificación de los rayos en la imagen capturada se utilizan diferentes métodos o técnicas del tipo multivariables que son abordadas en la tercera sección del presente capitulo.

## 3.1 DESCRIPCIÓN DE LA BASE DE DATOS

La principal herramienta para obtener información del problema es la respectiva base de datos, en la cual sus características generales son:

- Título: MAGIC gamma telescope data 2004
- Dueño original de la base de datos: R. K. Bock Major Atmospheric Gamma Imaging Cherenkov Telescope project (MAGIC)
- Donante: P. Savicky Institute of Computer Science, AS of CR Czech Republic
- Fecha: Mayo, 2007
- Usos pasados: Rule Induction in Forensic Science
    - Bock, R.K., Chilingarian, A., Gaug, M., Hakl, F., Hengstebeck, T., Jirina, M., Klaschka, J., Kotrc, E., Savicky, P., Towers, S., Vaicilius, A., Wittek W. (2004). Methods for multidimensional event classification: a case study using images from a Cherenkov gamma-ray telescope. Nucl.Instr.Meth. A, 516, pp. 511-528.
    - P. Savicky, E. Kotrc. Experimental Study of Leaf Confidences for Random Forest. Proceedings of COMPSTAT 2004, In: Computational Statistics. (Ed.: Antoch J.) - Heidelberg, Physica Verlag 2004, pp. 1767-1774.
    - J. Dvorak, P. Savicky. Softening Splits in Decision Trees Using Simulated Annealing. Proceedings of ICANNGA 2007, Warsaw, (Ed.: Beliczynski et. al), Part I, LNCS 4431, pp. 721-729.
- Información relevante: Los datos son generados para simular el registro de partículas gamma de alta energía en un telescopio Cherenkov utilizando la técnica de imagen. El Telescopio Cherenkov gamma observa rayos gamma de alta energía, tomando ventaja de la radiación emitida por las partículas cargadas producidas en el interior de las lluvias electromagnéticas iniciadas por los gammas en la atmósfera. Esta radiación Cherenkov (visible a longitudes de onda UV) se filtra a través de la atmósfera y

se registra en el detector, lo que permite la reconstrucción de los parámetros de la lluvia. La información disponible se compone de pulsos dejados por los fotones Cherenkov entrantes en los tubos foto-multiplicadores, dispuestos en un plano a la cámara. Dependiendo de la energía de la gamma primaria, un total de unos pocos cientos a unos 10.000 fotones Cherenkov se logran recoger, haciendo posible discriminar estadísticamente aquellas que son causadas por gammas primarias (señales) de las imágenes de lluvias hadrónicas iniciados por los rayos cósmicos en la atmósfera superior.

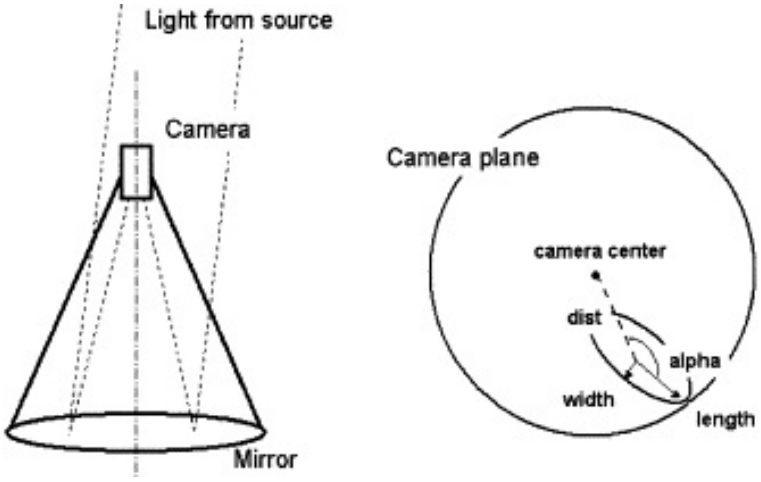


*Figura 3-1: Telescopio Cherenkov: sketch y definición de algunos parámetros de imágenes.*

El dataset completo contiene 2 archivos diferentes, los que se describen a continuación:

**magic04.data:** Archivo de principal predilección de la base de datos, que contiene los valores de cada partícula de interés (gamma, hadrónico). Los datos que contiene se presentan más adelante.

**glass.names:** Archivo que contiene información asociada a la base de datos, como el nombre del creador, de o los donante/s, los usos que se le ha dado anteriormente, etc. Esta información ha sido descrita anteriormente en esta sección.

La base de datos contiene 11 atributos por partícula, los cuales se describen a continuación:

1. fLength: Mayor eje de la elipse [mm].
2. fWidth: Menor eje de la elipse [mm].
3. fSize: 10-log de la sima de los contenidos de todos los píxeles [en la imagen].
4. fConc: Proporción de la mayor suma de dos píxeles sobre fSize [proporción].
5. fConc1: Proporción del mayor píxel sobre fSize [proporción].
6. fAsym: Distancia desde el mayor píxel al centro, proyectado sobre el eje mayor [mm].

7. fM3Long: Tercera raíz del tercer momento a lo largo del eje mayor [mm].
8. fM3Trans: Tercera raíz del tercer momento a lo largo del eje menor [mm].
9. fAlpha: Ángulo del mayor eje con vector al origen [grados].
10. fDist: Distancia desde el origen al centro de la elipse[mm].
11. class: Naturaleza gamma o hadrónico (g,h).

La base de datos tiene un total de 19020 instancias, en donde 12332 son de tipo gamma y 6688 de tipo hadrónico. Además como se menciono antes cuentan con los 10 atributos mencionados y un atributo de clase, que es el tipo (gamma, hadrónico). Es importante mencionar que todos los atributos son evaluados de manera continua.

## 3.2 ESTADO DEL ARTE

A continuación se describe los métodos multivariantes de clasificación con el fin de determinar (mediante una comparación) cuál de ellos es más efectivo al aplicarlos en el dataset.

### 3.2.1 Selección directa en los parámetros de la imagen

Se seleccionan por parámetros de corte en el caso unidimensional, también se puede aplicar en el n-espacio (en este caso los parámetros de imagen). Este es un método comúnmente utilizado entre los físicos. Cualquier método que pretenda ser superior, debe utilizar los resultados de estos como vara de medición. El método de selección directa también hace necesario que se establezca un criterio de optimización, y normalmente no dará lugar a una relación entre la aceptación gamma y la aceptación de hadrónes, aunque el plano de aceptación se rellenará mediante la variación de los criterios de selección .

### 3.2.2 Bosques y arboles de clasificación

Para la clasificación, se tiene un árbol de decisión binaria, en los cuales sus nodos internos representan pruebas comparándose un predictor por nodo con un umbral fijo. Cada una de las hojas está marcada por una de las dos clases: señal o fondo. El árbol puede construirse a partir del set de entrenamiento utilizando diferentes estrategias brevemente descritas más adelante. La clasificación de un nuevo caso se inicia en el nodo raíz del árbol. Se evalúa la prueba asignada a este nodo y el cálculo continúa el subárbol izquierdo o derecho de acuerdo con la respuesta. Esto se realiza repetidamente hasta que se alcanza una hoja. Su etiqueta representa la predicción resultante.

Una mejora significativa en la exactitud de un solo árbol de clasificación se puede lograr mediante la construcción de una colección de varios árboles diferentes (un bosque o un conjunto de predictores), usando además una votación o un proceso de promedio para la clasificación de nuevos casos. La construcción de un conjunto de predictores puede mejorar la precisión, no sólo de los árboles de decisión, sino también otros tipos de predictores, cuando se trata de predictores individuales de sesgo bajo pero de gran varianza.

Los experimentos alcanzados en la investigación, utilizaron tres estrategias diferentes para la construcción de los bosques. Los dos primeros utilizaron dos técnicas clásicas, CART (versión 4.0, que utiliza una

muestra boostrap para el cultivo del árbol y todo el conjunto de datos para la poda) y C4.5 / C5.0 (release 1.15, que que utiliza el algoritmo C4.5 para la inducción de árboles individuales, permitiendo la opción de usar los bosques), para la construcción de varios árboles, que fueron combinados para formar un bosque. Con el fin de obtener diferentes árboles utilizando C5.0 y CART, se utilizó una técnica de ”bagging”(agregación bootstrap), donde cada árbol se cultiva en una sub-muestra aleatoria de los datos de entrenamiento, dibujadas con o sin sustitución. El último experimento utilizó la técnica de bosque aleatorio, que está directamente diseñado para construir un bosque.

#### *3.2.2.1 Random Forest*

En el método aleatorio forestal, una vez más un gran número de árboles de clasificación se cultiva y se combina. Dos elementos aleatorios sirven para obtener un bosque al azar, embolsado y selección dividida al azar.

El bagging es hecho en esta parte mediante el muestreo varias veces con el reemplazo del conjunto de datos de entrenamiento originales. Así, en las muestras resultantes, un determinado evento puede aparecer varias veces, y otros eventos no totalmente. Aproximadamente 2/3 de los datos en la muestra de entrenamiento se toman para cada muestra bootstrap.

La selección aleatoria de división se utiliza en el proceso de crecimiento de cada árbol. El cultivo de árboles comienza con todos los casos que se encuentran en el nodo raíz. El nodo raíz se divide entonces por un corte usando uno de los parámetros de la imagen, en dos nodos sucesivos para lograr una clasificación por separación de las clases.

### 3.2.3 Métodos kernel

El Kernel o núcleo de probabilidad de densidad de Estimación (PDE) se puede utilizar como una técnica no paramétrica de clasificación multivariante, y se basan en la premisa de que la función de densidad de la distribución que es desconocida, se puede aproximar por una suma de algún otro elegido entre las funciones kernel.

Hemos restringido nuestros estudios a los métodos de PDE basados en un núcleo de Gauss, que es una elección natural para la mayoría de aplicaciones de análisis de la física, ya que casi todas las variables utilizadas en el análisis han sido por lo general Gaussianas, manchadas por la resolución del detector u otros efectos.

Una aplicación típica de PDEs gaussiana comienza con una muestra de n eventos de Monte Carlo generados en un espacio de parámetros k-dimensional. Los eventos de Monte Carlo se distribuyen de acuerdo a alguna función de densidad (desconocida) de probabilidad (PDF).

Métodos Kernel PDE, en general, hacen un excelente trabajo de modelar complejas interrelaciones en espacios de alta dimensión de parámetros, siempre y cuando los PDF están variando sin problemas dentro de ese espacio de parámetros. De este modo, a menudo se desempeñan tan bien o incluso mejor que las técnicas más sofisticadas, como los árboles de clasificación o redes neuronales artificiales.

### 3.2.4 Redes neuronales artificiales (ANN-s)

Los métodos ANN-s se parecen a los métodos basados en los árboles en que se definen las selecciones simultáneas en las variables, pero en lugar de las variables originales, se trabaja con los datos a nivel local linealmente transformados, y la propia transformación es parte del proceso de optimización. Por lo general, los resultados coinciden, pero no son de calidad superior a los resultados de referencia que existen para la comparación.

#### *3.2.4.2 Paquete NeuNet*

Este paquete permite la construcción de múltiples capas de redes de alimentación. Otras características de este paquete de red neuronal son la inicialización aleatoria de pesos y sesgos.

#### *3.2.4.3 NNSU- Red neuronal con unidades de conmutación*

El NNSU es una combinación de una arquitectura de red neural clásica y un árbol de clasificación. Esta red es en realidad un gráfico acíclico orientado cuyos nodos son estructuras llamadas bloques de construcción. Este gráfico acíclico es referido con una gráfica exterior. Cada bloque de construcción es una red neural que consiste en dos tipos de nodos. Estos nodos están conectados de tal manera que forman un grafo acíclico de nuevo, con la restricción de que la dimensión de salida de cada bloque de construcción es el mismo para todos los bloques de construcción en la gráfica exterior. El primer tipo de nodo, al que llamamos unidad funcional, hace un mapeo predefinido en el espacio de entrada correspondiente a este nodo para el espacio de salida de este mismo. Por lo tanto un nodo de este tipo puede ser descrito por una tupla de números enteros, un vector de entrada y un vector de salida, y por una función de transferencia. La definición de esta función de transferencia incluye los parámetros de esta unidad funcional (vectores de peso, etc. umbral en la terminología actual de la red neural).

#### *3.2.4.4 GMDH—Método de Manejo de grupo de datos*

El Group Method Data Handling (método de manejo de grupos de datos), es un aproximador polinómico, cuyo grado del polinomio es controlado de acuerdo con la calidad de la aproximación alcanzada. La selección de un grado apropiado de polinomio es un problema bien conocido de aproximación polinómica. Debido a que el grado del polinomio se controla inherentemente según el carácter de la tarea resuelta, también el tamaño de las correspondientes redes neuronales, es controlado. La descripción del algoritmo GMDH se cita de la siguiente manera, se empieza calculando las ecuaciones de regresión para cada par de variables de entrada y de salida. Esto da las variables de orden superior para aproximar la salida en lugar de las variables originales. Después de encontrar estas ecuaciones de regresión (de un conjunto de observaciones de entrada / salida), se da cuenta de cuál guardar. Esto da una colección de modelos de regresión de segundo grado, que mejor aproxima (tenga en cuenta que cada aproximación depende de dos variables independientes).

Después del proceso anterior, se usan cada una de las ecuaciones de segundo grado obtenidas y se generan nuevas observaciones independientes (que sustituyen a las observaciones originales de las variables). A partir de estas nuevas variables independientes se obtienen combinaciones como se hizo en el proceso anterior. Es decir, se calculan todas las ecuaciones de regresión cuadrática de la salida frente a estas nuevas

variables. Esto da una nueva colección de ecuaciones de regresión para la aproximación de la salida para las nuevas variables, que a su vez son estimaciones de la salida para las ecuaciones anteriores. Después se selecciona lo mejor de estas estimaciones, se generan nuevas variables independientes de las ecuaciones seleccionadas para reemplazar las anteriores, y se combinan todos los pares de estas nuevas variables.

El proceso continua hasta que se alcance algún criterio de convergencia. Cada ecuación de regresión surge de la combinación de dos variables del conjunto anterior. Uno puede ver el proceso como la construcción de la red neuronal mediante la adición de dos neuronas de entrada a la anterior. En lugar de la función sumatoria estándar y una función de transferencia, se utiliza la función cuadrática anterior.

#### *3.2.4.5 MRS y MLP redes neuronales*

Se incluyen dos análisis más con redes neuronales. Una red de alimentación con MultiStart búsqueda aleatoria (MRS), y un ajuste perceptrón multicapa (MLP). El enfoque MRS utiliza un subconjunto de parámetros considerados óptima, y selecciona múltiples pesos iniciales y configuraciones de red, de la que conserva el que tiene el mejor resultado (en la muestra de control). El MLP es parte del paquete Terraferma y pone el acento en los métodos de entrenamiento de gran alcance y el cálculo de doble precisión.

### 3.2.5 Vecino más cercano

Como se vio anteriormente, los métodos kernel asignan ponderaciones a los sucesos en la muestra de referencia, que se desvanecen rápidamente si la distancia desde el nuevo punto supera un cierto radio estipulado por el parámetro h. Un efecto similar se puede llegar aún más directamente por mirar sólo a los eventos en una ventana de un radio determinado. Sin embargo, la densidad de puntos en diferentes regiones del espacio de parámetros es diferente. Esto conlleva a intentar también un método que utiliza un número fijo de vecinos no ponderados más cercanas. De hecho, los métodos del vecino más cercano son simples, pero su problema es el de definir una métrica válida.

Una posible reducción del espacio de parámetros se ha explorado también con este método, aunque superficialmente: sólo hay una pequeña pérdida en la calidad de clasificación cuando se reduce de diez a ocho parámetros, pero más allá de eso, las pérdidas se hacen visibles. La elección de qué parámetros se quedan fuera no parece ser relevante.

El método del vecino más próximo requiere como única opción la de una muestra de referencia (como el método de kernel), además del número de vecinos más cercanos a considerar. Se han utilizado las muestras de formación, definidas para todos los otros métodos, pero los resultados no muestran cambios si se utilizan muestras de referencia inferior a los 1500 eventos, o si las muestras de control y de formación son las mismas.

### 3.2.6 Soporte de Maquinas de vectores

Los soportes de maquinas de vectores (SVM-s) se encuentran actualmente en una investigación muy activa dentro de los campos de la computación neuronal y aprendizaje automático. SVMs son ejemplos de una categoría más amplia de enfoques de aprendizaje que utilizan el concepto de sustitución del kernel, con lo que la tarea de aprender es más dúctil mediante la explotación de una asignación implícita en un

espacio dimensional alto. Motivado por la teoría del aprendizaje estadístico que se han aplicado con éxito a numerosas tareas dentro de la minería de datos, la visión por computador y la bioinformática.

### 3.2.7 Probabilidades compuestas

Este método no publicado, utiliza probabilidades de eventos obtenidos mediante la comparación de los datos del evento de densidades de probabilidad bidimensionales obtenidas a partir de una muestra de entrenamiento. Las densidades son determinados por la histografía de los datos de entrenamiento en dos dimensiones, utilizando contenedores que dan contenido a los comportamientos constantes de datos de la señal. Todas las proyecciones 2D se utilizan que se puede hacer a partir de los parámetros de la imagen, es decir, en el caso del estudio, con diez parámetros 45 proyecciones. Cada comportamiento de cada proyección termina con una probabilidad de ser la señal (debido a la definición de comportamiento siempre cerca de los mismos, 1 / nbins), y una probabilidad de ser de fondo. Por tanto, un evento de entrada tiene que ser desechado; las probabilidades asociadas con la muestra de entrenamiento para los contenedores en cada proyección se multiplican, y el producto, la probabilidad compuesta, se toma como una sola estadística de prueba señalitica.

### 3.2.8 Análisis discriminante lineal (LDA)

Este es un método popular sobre todo porque da lugar a un cálculo paramétrico elegante. Su objetivo es encontrar una combinación lineal de los parámetros originales de la imagen de tal manera que el hiperplano definido por la transformación maximice la distancia entre los medios las muestras de señal (gamma) y de fondo (hadrónes), minimizando al mismo tiempo la varianza dentro de cada muestra. El método es rápido, simple y robusto; también no depende de muestras de entrenamiento. Sin embargo, no tiene en cuenta las correlaciones no lineales en el espacio n-dimensional (a causa de la transformación lineal). Los resultados inferiores que se logran con LDA indican que por lo menos las variables de orden superior deben introducirse. Hay variantes a LDA como Análisis Discriminante Cuadrático (QDA) y Análisis Discriminante Regularizado (RDA), que responde en parte a esta crítica.

LDA se utiliza para la reducción de dimensionalidad: LDA reduce el espacio de alta dimensión de una sola variable de mejor separación; PCA, reduce un espacio de alta dimensión a un espacio en el que no todas las variables tienen el mismo significado, lo que permite hacer caso omiso de algunos de ellos (regularización). La diferencia entre LDA y PCA, es que PCA realiza la clasificación de elementos (por ejemplo, algunos de los parámetros de la imagen a partir de nuestros datos telescopio Cherenkov se obtienen mediante PCA), mientras LDA realiza la clasificación de la muestra.

### 3.2.9 Comparación de los diferentes métodos

A continuación se muestra una tabla con el resumen de los resultados obtenidos al aplicar los diferentes métodos descritos a la muestra (proveniente del dataset de la base de datos).

*Tabla 3.1: Resumen de medidas estadísticas*

| Método | loacc | hiacc | Q(0.5) | $\sigma$ (0.5) | $\sigma$ (max) | $\varepsilon$ (y) |
|---|---|---|---|---|---|---|
| Random Forest | 0.452 | 0.852 | 2.8 | 8.44 | 8.74 | 0.412 |
| C5.0 | 0.441 | 0.830 | 2.7 | 8.14 | 8.96 | 0.408 |
| CART | 0.414 | 0.810 | 2.6 | 7.94 | 8.03 | 0.538 |
| Vecino Cercano | 0.448 | 0.816 | 2.6 | 8.03 | 9.12 | 0.317 |
| Kernel | 0.443 | 0.803 | 2.8 | 8.43 | 8.64 | 0.390 |
| NNSU | 0.472 | 0.731 | 3.5 | 9.74 | 9.82 | 0.483 |
| NeuNet | 0.445 | 0.839 | 3.0 | 8.73 | 8.75 | 0.483 |
| MRS | 0.348 | 0.779 | 2.3 | 7.16 | 7.31 | 0.431 |
| MLP | 0.300 | 0.767 | 2.2 | 6.93 | 7.22 | 0.576 |
| GMDH | 0.280 | 0.736 | 2.0 | 6.55 | 6.77 | 0.574 |
| Prob. compuestas | 0.332 | 0.728 | 2.1 | 6.78 | 6.83 | 0.585 |
| Selección Directa | 0.306 | 0.636 | 1.8 | 5.91 | 7.52 | 0.153 |
| LDA | 0.195 | 0.638 | 1.6 | 5.47 | 5.80 | 0.710 |
| SVM | 0.124 | 0.586 | 1.4 | 4.81 | 5.76 | 0.784 |

Como se puede apreciar se utilizan 6 parámetros de calidad,entre los cuales se encuentran loacc, hiacc, significancia $\sigma$ y el factor de calidad Q. En donde **loacc** es la media aritmética de los valores de señal (gamma) y se obtiene mediante la interpolación de la curva en los puntos 0.01, 0.02, 0.05. Por otro lado **hiacc** es obtenido de manera similar al promediar $\epsilon$ (gamma) en los puntos 0.1 y 0.2.

Para tener una visión mas gráfica de los resultados puede verse el gráfico que se muestra a continuación.

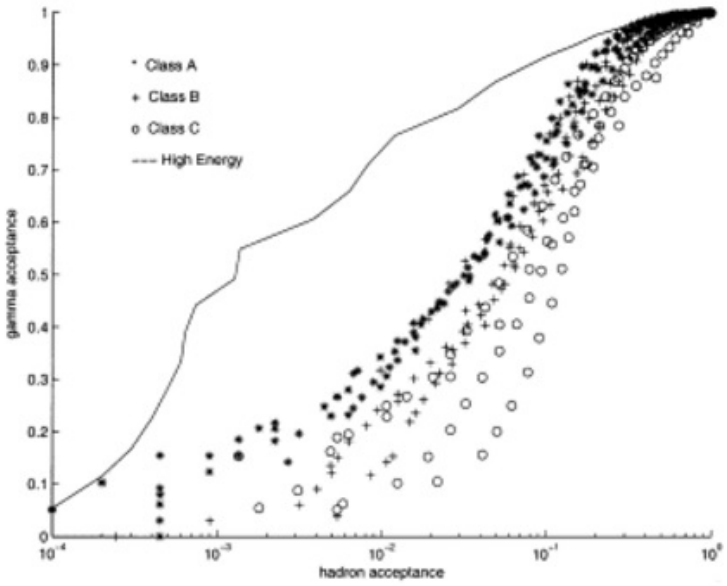


*Figura 3-2: Comparación de los distintos métodos de clasificación.*

Como se puede apreciar en el gráfico, los distintos métodos fueron divididos en clase A que son los arboles de regresión, clase B que contiene los métodos de red neuronal, clase C que es SVM (soporte de maquinas de vectores), selección directa, LDA y las probabilidades compuestas. Por ultimo para energías altas que hace referencia al método del vecino mas cercano.

# CAPÍTULO 4. CONCLUSIONES

Se puede observar que de acuerdo a la comparación de los métodos hecha en la sección anterior, que los resultados de árboles de clasificación, kernel y el método del vecino más cercano tienen similitudes. En cuanto al (ANN) o métodos de redes neuronales, se obtienen resultados en un amplio rango, entre los mejores y de resultados medios. Esto da a entender, que los métodos ANN necesitan una comprensión más detallada, y que sin duda no se pueden utilizar fuera de la plataforma. Y para el último grupo de métodos (probabilidades compuestas, selección directa, LDA y SVM), se puede decir que estos muestran ser inferiores en cuanto a sus resultados.

Hablando de los tres métodos de clasificación de árboles, estos utilizan varios árboles, en el caso de los árboles individuales, los resultados son inferiores y no se muestran. El método Random Forest superó claramente a los resultados obtenidos con las metodologías C5.0 y CART.

Las conclusiones que se pueden extraer de los resultados obtenidos son válidos para los datos de entrada de este estudio en particular. Un nuevo uso de esas conclusiones a diferentes muestras de datos deben ser validadas nuevamente.

Los métodos en el estudio asumen un espacio abstracto de los parámetros de la imagen, que se adapta bien a las situaciones de los eventos de Monte Carlo y tal vez sólo a estos. Para este estudio no se consideraron las posibles influencias que pueden distorsionar este espacio. En el caso de los telescopios Cherenkov, el campo de estrellas en el campo de visión y el cielo nocturno, como también el cambio de fondo durante la observación, además de la constante variación de las condiciones atmosféricas, pueden producir cambios en el detector que son inevitables, provocando posiblemente un mal funcionamiento. Algunos de estos efectos observacionales son atendidos por la transformación de los contenidos de píxeles a los parámetros, es decir, pre-procesamiento.

Ningún método de clasificación por sí solo, puede sustituir las mejoras de pre-procesamiento de la imagen, o, para el caso, inventar nuevos parámetros independientes que contengan más información. Ellos se pueden derivar de la imagen, pero también se podrían deducir de forma independiente de las observaciones, por ejemplo, la hora de llegada o la energía de los fotones, para encontrar uno de estos, se requiere de algunos conocimientos físicos y la buena comprensión del detector.

Finalmente con respecto a los objetivos inicialmente planteados, se ha logrado adquirir una idea inicial acerca del tema de investigación, por ende, para posteriores interpretaciones e inferencias acerca de los datos aquí estudiados, se tiene una base firme que permitirá desarrollar futuras investigaciones.

Con respecto a los métodos, tal como se dijo anteriormente, se ha logrado dilucidar que técnicas son más precisas para el respectivo dataset.

# CAPÍTULO 5. BIBLIOGRAFÍA