Métodos de Aprendizaje Computacional para la Detección Automática de Afectaciones Ionosféricas en Sistemas de Aumentación Terrestres

Wanda Catalina Rincón Cadena

Trabajo de Grado para Optar el título de

Magister en Matemática Aplicada

Director

Ph.D. Raúl Ramos Pollán

Universidad Industrial de Santander

Facultad de Ciencias

Escuela de Física

Maestría en Matemática Aplicada

Bucaramanga

2018

Índice

RESUMEN

TÍTULO: MÉTODOS DE APRENDIZAJE COMPUTACIONAL PARA LA DETECCIÓN AUTOMÁTICA DE AFECTACIONES IONOSFÉRICAS EN SISTEMAS DE AUMENTACIÓN TERRESTRES.

AUTOR: WANDA CATALINA RINCÓN CADENA.

PALABRAS CLAVE: APRENDIZAJE COMPUTACIONAL, ANOMALÍAS IONOSFÉRICAS, GNSS, GBAS, ANÁLITICA DE DATOS.

DESCRIPCIÓN:

Los usuarios de aviación civil necesitan de los sistemas de posicionamiento global como GPS para dar soporte a sus operaciones. Las anomalías ionosféricas son una de las principales amenazas para la seguridad y disponibilidad de las medidas de posición que provee el sistema GPS. La ionósfera es una capa de la atmósfera que se encuentra cargada con electrones, estos afectan todas las comunicaciones por ondas de radio, debido a que refractan parte de las señales que atraviesan este medio. A través de una metodología para corregir las señales de código y fase captadas por receptores de doble frecuencia, se calculan y preprocesan los retrasos ionosféricos en las señales de pseudorango, seguido del cálculo de gradientes entre pares de receptores cercanos alrededor de un aeropuerto para estimar el comportamiento de la ionosfera en el área. Sin embargo, después de este proceso existen falsos positivos entre los resultados. En este trabajo se propone como solución la creación de conjuntos de datos y el diseño de características de eventos ionosféricos para entrenamiento de modelos de aprendizaje computacional, que clasifiquen cada caso como verdadero o falso. De esta manera, se provee una solución de bajo costo para evaluar la ionósfera en una región previo a la instalación de estaciones de monitoreo. Con los datos recolectados se construye un conjunto de descriptores de las señales captadas por los receptores además de información de clima espacial como actividad solar y geomagnética. Los receptores utilizados para construir el conjunto de datos pertenecen a redes de Ecuador, Estados Unidos y España. Los resultados sugieren que es posible automatizar la validación de eventos ionosféricos extremos con un desempeño en la métrica f2 de hasta 93 %.

* Master Thesis
** Facultad De Ciencias Básicas. Escuela de Física.
Maestría en Matemática Aplicada. Director: Raul Ramos Pollán.

ABSTRACT

TITLE: MACHINE LEARNING METHODS FOR IONOSPHERE ANOMALY DETECTION ON GROUND BASED AUGMENTATION SYSTEMS.

AUTHOR: WANDA CATALINA RINCÓN CADENA.

KEYWORDS:MACHINE LEARNING, IONOSPHERE ANOMALY, GNSS, GBAS, DATA ANALYTICS.

DESCRIPTION:

Civil aviation users need global positioning systems such as GPS to support their operations. Ionospheric anomalies are one of the main threats to security and availability of position measurements provided by GPS, the ionosphere is the atmosphere layer charged with electrons that affects all communications through radio waves. Using a methodology, ionospheric delays are calculated and preprocessed in pseudorange signals from double frequency receivers that are located in the area around an airport, then gradients are calculated between pairs of nearby receivers to estimate the behavior of the ionosphere. However, after this process there are false positives among the results. In this work we proposed as a solution machine learning models to classify each case as true or false based on the input characteristics, in this way, a low cost solution can be provided to evaluate the ionosphere before the installation of expensive equipment. From the data collected, a database is constructed with information of the signals received by the sensors, as well as spatial weather information such as solar and geomagnetic activity. The receivers used to build the dataset belong to networks in Ecuador, the United States and Spain. The results suggest that it is possible to automate the validation of extreme ionospheric events with a performance in the f2 metric of 93 %.

* Master Thesis
** Facultad De Ciencias Básicas. Escuela de Física.
Maestría en Matemática Aplicada. Director: Raul Ramos Pollán.

1. Introducción

Los sistemas globales de navegación por satélite (GNSS) proveen a sus usuarios de una estimación de su posición. Muchas aplicaciones civiles necesitan de los sistemas GNSS para su funcionamiento, entre estas: el apoyo de operaciones de aviación, sistemas inteligentes de transporte terrestre y seguimiento de mercancía. Las señales GNSS al ser enviadas por el satélite, deben viajar por el espacio y atravesar la atmósfera terrestre antes de llegar al receptor. Esto supone afectaciones en la señal enviada: ruido, error de multicaminos y retrasos causados por la naturaleza de la atmósfera son sólo algunos de los inconvenientes que enfrentan los GNSS.

La mayor fuente de imprecisión en la región ecuatorial y latitudes bajas es el retraso de la señal causado por el comportamiento de la ionósfera. La ionósfera es un medio muy dispersivo e irregular, en especial en ciclos solares fuertes. Existen varias propuestas para modelar e identificar su comportamiento de diferentes maneras como en [7] y [4]. La aviación requiere de aplicaciones de alta precisión, para que un avión pueda aterrizar con seguridad se emplean sistemas GBAS (Ground Based Augmentation Systems) que con ayuda de una red de estaciones terrestres cercanas a la pista de aterrizaje (menos de 100 kilómetros), estos sistemas monitorean y proveen alertas de estados peligrosos para aumentar la integridad de un GNSS. La ionósfera causa dos inconvenientes para los sistemas GBAS: Gradientes espaciales y centelleo. Los gradientes son el retraso en la recepción de la señal, el cual se ve reflejado en una gran variación del error en posición. Mientras que el centelleo es la fluctuación en la intensidad y fase de la señal portadora, debido a la distribución espacial irregular de los electrones en la ionósfera.

El estudio y procesamiento de datos de gradientes en GBAS es aún semi-automático [6], pues con la ayuda del monitoreo que hacen profesionales expertos en el tema se valida si la imprecisión de las posiciones es causada por un frente ionosférico, o por otras causas como por ejemplo fallas en el receptor. Este trabajo plantea una solución automática para la clasificación de eventos ionosféricos en GBAS, usando la información proveniente de los retrasos de señales procesadas, junto con productos de clima espacial disponibles para el público en las bases de datos de la Administración Nacional Oceánica y Atmosférica de los Estados Unidos (NOAA) para entrenar algoritmos de aprendizaje computacional en la tarea de clasificar cuando un gradiente observado es una verdadera amenaza. Para poder tener medidas del gradientes se desarrolló la metodología [14]. En nuestro flujo de trabajo de aprendizaje estadístico se diseñaron experimentos para entrenar y probar el desempeño de diferentes algoritmos de clasificación hasta llegar a los modelos finales que muestran resultados satisfactorios.

---

* Master Thesis

** Facultad De Ciencias Básicas. Escuela de Física.
Maestría en Matemática Aplicada. Director: Raul Ramos Pollán.

2. Objetivos

2.1. Objetivo General

Desarrollar métodos basados en modelos matemáticos de aprendizaje computacional que detecten anomalías ionosféricas para algoritmos de navegación basados en GNSS diferencial aplicados en contextos GBAS.

2.2. Objetivos Específicos

- Desarrollar la algorítmica de procesado de datos GNSS para la validación de riesgos de amenazas ionosféricas de GBAS, según los métodos semiautomáticos existentes en el estado del arte [14].
- Desarrollar métodos basados en aprendizaje computacional para la automatización completa de la validación de riesgos de amenazas ionosféricas de GBAS.
- Validar los métodos desarrollados según las métricas estándar de performance en navegación aérea.
- Valorar la factibilidad de implantación de los métodos desarrollados en un entorno de producción en función de sus demandas computacionales.

* Master Thesis
** Facultad De Ciencias Básicas. Escuela de Física.
Maestría en Matemática Aplicada. Director: Raul Ramos Pollán.

3. Marco teórico

Para comprender mejor el proyecto se hace un resumen acerca de los métodos de aprendizaje computacional y por ser el conjunto de datos de proveniente del sistema GPS, un breve resumen del funcionamiento de los sistemas de posicionamiento satelital y sus fuentes de error.

3.1. Algoritmos de Aprendizaje Computacional

Durante la última década la humanidad ha generado más de 90 % de los datos del mundo [16], esto gracias a los avances tecnológicos que han causado el fenómeno de "Big data". Se ha podido generar, guardar y procesar más información que nunca antes. Aprender nuevo conocimiento a partir de esta gran cantidad de datos es posible a través de las técnicas del campo de aprendizaje computacional también llamado aprendizaje estadístico. El aprendizaje computacional comprende diversas técnicas que son como un sistema de inducción, su objetivo es generar modelos que permitan generalizar a partir de los datos conocidos, encontrando patrones que permitan hacer predicciones en nuevos datos.

El aprendizaje computacional permite generar soluciones que tardarían mucho tiempo en ser generadas de manera "manual" modelando ciertos fenómenos, hoy en día estas técnicas se encuentran implementadas en numerosas aplicaciones de la vida cotidiana como lo son: sistemas de recomendación de compras, reconocimiento de imágenes y vehículos autónomos, entre otras. Los algoritmos se pueden clasificar según el tipo de aprendizaje, esto es si un aprendiz recibe ejemplos etiquetados con las respuestas que se quieren obtener entonces el aprendizaje es supervisado de lo contrario es no supervisado.

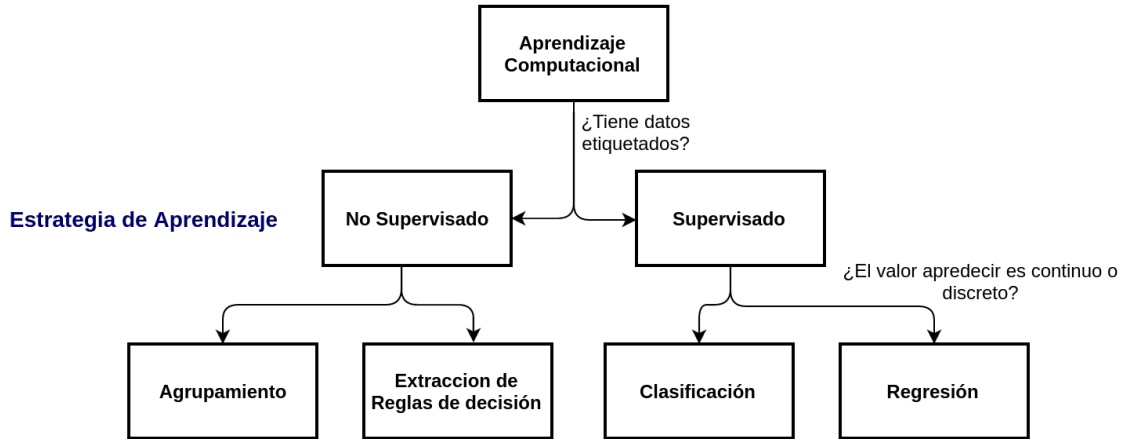

Figura 1: Clasificación de algoritmos de aprendizaje computacional.

De la misma manera podemos clasificar los modelos según el tipo de variable

* Master Thesis
** Facultad De Ciencias Básicas. Escuela de Física.
Maestría en Matemática Aplicada. Director: Raul Ramos Pollán.

que queremos obtener, podemos plantear un modelo de aprendizaje computacional cuya entrada es una matriz X de dimension m por n y uno de los ejemplos de entrenamiento $X_i = (x_1, x_2, ...x_j, ..., x_n)$ con n características y el vector de etiquetas o variable respuesta de tamaño m por 1. La meta es generar un modelo que genere los parámetros tal que para todo minimizando el error o diferencia entre y_i y $\hat{y}_i$, que se cuantifica mediante una función de costo. Si tenemos que y_i es una variable discreta entonces estamos tratando un problema de clasificación, el objetivo será clasificar un elemento X_i en las clases o valores posibles de y_i. De lo contrario si y_i es una variable continua estaríamos usando técnicas de regresión. Este trabajo propone validar si un gradiente identificado por la estación GBAS es verdadero o falso, además contamos con eventos etiquetados. Por lo cual nos centraremos en las técnicas de aprendizaje supervizado para clasificación.

3.1.1. Modelos Lineales de Clasificación

Los modelos de clasificación lineal como su nombre lo indica buscan crear fronteras o límites de decisión lineales. Para K clases el modelo lineal de la k-ésima variable respuesta es $\hat{f}_k(x) = \hat{\beta}_{k0} + \hat{\beta}_k^T x$ Para cualquier par de clases k y l el límite de decision será aquellos puntos donde $\hat{f}_k(x) = \hat{f}_l(x)$ definiendo un hiperplano, el espacio de entrada será dividio en regiones de constante clasificación. También pertenecen a esta clase de métodos aquellos que modelan funciones discriminantes para cada clase y luego clasifican X_i en la clase con el mayor valor en su función discriminante . Sólo requeriremos de una transformación monótona de $\delta_k(x)$ para que los límites sean lineales.

Análisis de Discriminante Lineal

La teoría de decisión nos dice que necesitamos saber la probabilidad de clase a posteriori $Pr(G|x)$ para una clasificación óptima. Suponga que $f_k(x)$ es la función de densidad condicinal de clase de X en la clase k , con la condición de$\sum_{k=1}^{K} \pi_k = 1$

Al aplicar el teorema de Bayes obtenemos:

$Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}\pi_l$

En este método se usan funciones de densidad de clase gaussianas, multivariadas:

$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}$ Donde p es el número de dimensiones.

El análisis de discriminante lineal es el caso especial donde asumimos que **las clases tienen una matriz de covarianza igual** $\Sigma_k = \Sigma \forall k$. Al comparar las dos clases k y l con el radio de logaritmos se ve que:

$log\frac{Pr(G=k|X=x)}{Pr(G=l|X=x)} = log\frac{f_k(x)}{f_l(x)} + log\frac{\pi_k}{\pi_l} = log\frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^T\Sigma^{-1}(\mu_k + \mu_l)$

---

* Master Thesis
** Facultad De Ciencias Básicas. Escuela de Física.
Maestría en Matemática Aplicada. Director: Raul Ramos Pollán.

El conjunto de puntos donde $Pr(G = k|X = x) = Pr(G = l|X = x)$ constituye el hiperplano separador de las clases k y l. En la práctica no conocemos los parámetros de las distribuciones, y estas se deben estimar de los datos de entrenamiento (muestra):

$\hat{\mu}_k = \sum_{g_i=k} x_i/N_k$ donde N_k es el tamaño de la muestra de clase k. $\hat{\Sigma}$ es la covarianza.

$\hat{\pi}_k = N_k/N$ probabilidad a priori de que un elemento pertenece a la clase k.

Regresión Logística

Nace del deseo de modelar probabilidades posteriores de las K clases con funciones lineales en x, asegurandose que las probabilidades sumen 1 y se mantengan en el rango [0,1]. En la regresión logística modelamos una probabilidad $p(X)$ de la siguiente manera:

$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$ Sin importar qué valores tome β o X, la propabilidad se mantendrá en el intervalo [0,1].

En el caso de la clasificación binaria, la función logit nos da una función lineal: $log\frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X$

Ahora para K clases el modelo tiene la forma: $log\frac{Pr(G=1|X=x)}{Pr(G=K|X=x)} = \beta_{10} + \beta_1^T x$

$log\frac{Pr(G=2|X=x)}{Pr(G=K|X=x)} = \beta_{20} + \beta_2^T x$

... $log\frac{Pr(G=K-1|X=x)}{Pr(G=K|X=x)} = \beta_{(K-1)0} + \beta_{K-1}^T x$

K - 1 funciones logit, aunque la elección de la última clase en el denominador para todas funciones es arbitraria se puede comprobar que suman 1.

$Pr(G = k|X = x) = \frac{exp(\beta_{k0}+\beta_k^T x)}{1+\sum_{l=1}^{K-1} exp(\beta_{l0}+\beta_l^T x)}$ con $k = 1, 2, 3, ..., K-1$

$Pr(G = K|X = x) = \frac{1}{1+\sum_{l=1}^{K-1} exp(\beta_{l0}+\beta_l^T x)}$

Queremos encontrar el conjunto de parámetros $\theta = \{\beta_{10}, \beta_1^T, ..., \beta_{(k-1)0}, \beta_{k-1}^T\}$, denotamos las probabilidades de que los datos pertenecen a la clase k como $p_k(x;\theta)$. El modelo de regresión logística es ajustado por medio del método de máxima verosimilitud, usando la función de verosimilitud de G dado un X. El logarimo de verosimilitud para N observaciones es:

$\ell(\theta) = \sum_{i=1}^{N} logp_k(x_i;\theta)$

Donde $p_k(x_i;\theta)$ es la probabilidad de que x_i pertenezca al k usando los parámetros

θ. Para el caso de dos únicas clases y_i, los algoritmos se simplifican y se codifican las dos clases como 0 ó 1. Sea $p_1(x;\theta) = p(x;\theta)$ y $p21(x;\theta) = 1 - p(x;\theta)$. El logaritmo de la verosimilitud puede ser escrito como:

$\ell(\beta) = \sum_{i=1}^{N}\{y_i log p(x_i;\beta) + (1-y_i)log(1-p(x_i;\beta))\}$
$= \sum_{i=1}^{N}\{y_i\beta^T x_i - log(1+e^{\beta^T x_i})\}$

Para maximizar el logaritmo de la verosimilitud, se deriva e iguala a cero.

$\frac{\partial\ell(\beta)}{\partial\beta} = \sum_{i=1}^{N} x_i(y_i - p(x_i;\beta)) = 0$

Se resuelven las p+1 ecuaciones anteriores con el método de Newton-Raphson.

3.1.2. Modelos no Lineales de Clasificación

Clasificador Bayesiano Ingenuo

Similar a los métodos de clasificación lineal como regresión logística y análisis de discriminate lineal, comparte ventajas y desventajas con estos al ser rápidos de entrenar pero de bajo poder de generalización. Suele ser aplicado en grandes conjuntos de datos, y de gran número de características p. El clasificador de Bayes ingenuo asume que para cada clase las características son **independientes**. Por lo que se puede modelar la función de densidad de probabilidad de las características como: $f_j(X) = \prod_{k=1}^{p} f_{jk}(X_k)$, siendo j la clase y k una característica del conjunto de datos.

Aunque esta suposición no sea correcta en la mayoría de los casos, simplifica la estimación de las probabilidades, independiente del tipo de variable (continua o discreta). Similar a los métodos anteriores podemos derivar funciones logit:

$log\frac{Pr(G=\ell|X)}{Pr(G=J|X)} = log\frac{\pi_\ell f_\ell(X)}{\pi_J f_j(X)}$

$= log\frac{\pi_\ell}{\pi_J} + \sum_{k=1}^{p} log\frac{f_{\ell k}(X_k)}{f_{Jk}(X_k)}$
$= \alpha_\ell + \sum_{k=1}^{p} g_{\ell k}(X_k)$

Árboles de Decisión

Los árboles de decision son métodos que particionan el espacio de características en un conjunto de rectángulos y ajustan un modelo sencillo a cada característica. Una de las ventajas de este algoritmo es la transparencia que aporta, pues al final del proceso sabemos exactamente que características y cómo son usadas para hacer la clasificación de un elemento, esto aporta gran interpretabilidad y es una de las razones por las cuales los árboles de decisión siguen siendo usados hasta el día de hoy.

Un árbol clasificador consta de un conjunto de nodos, los nodos son una representación de particiones de características, habrá un nodo padre y varios nodos que

se derivan a partir de este junto con reglas de decisión que van dividiendo el espacio de entrada hasta que sólo quedan nodos "hoja". Cada nodo tiene un conjunto de ejemplos de entrenamiento y un valor dado por el criterio de impureza asociado a este. Esto es si un nodo cuenta con ejemplos que pertenecen a una única clase entonces el valor de impureza es 0. La impureza del nodo I se puede estimar con el criterio gini:

$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^2$ donde $p_{i,k}$ es el radio de instancias de la clase k en el nodo i.

En los árboles de decisión no hay optimización de una función convexa a diferencia de otros métodos como por ejemplo Regresión logística. El criterio se debe estimar para cada punto en cada característica con la intención de encontrar la mejor partición posible.

Otro de los criterios más usados es la entropía de un nodo, la cual se estima como:

$H_i = -\sum_{k=1,pi,k\neq 0}^{n} p_{i,k} log(p_{i,k})$

La entropía es un concepto que indica si un grupo esta desordenado, si su valor es cero entonces el nodo es homogéneo. La decisión de usar un criterio como la entropia o gini no suele afectar el desempeño de un clasificador.

Uno de los algoritmos de arboles más usados es el agoritmo CART (Classification and Regression Tree). En este, los datos son divididos en dos subconjuntos: izquierdo y derecho. Se busca el par (k, t_k) que de como resultado los subconjuntos más puros. La función de costo que se quiere minimizar es:

$J(k, t_k) = \frac{m_{left}}{m} G_{left} + \frac{m_{lright}}{m} G_{right}$

Una vez se ha dividido el conjunto de datos en dos con la mejor división, se aplica la misma regla a uno o ambos nodos, el proceso continúa hasta que una regla para detener el proceso sea aplicada, como una máxima profundidad permitida, o parámetros de impureza para los nodos hoja. Los árboles suelen ser métodos que se "sobreajustan" a los datos por lo cual se suelen aplicar mecanismos de "poda", esto consiste en identificar y remover nodos irrelevantes ya que estos tienen alta probabilidad de estar aprendiendo valores atípicos o errores de los datos. La complejidad de este algoritmo es de orden *O(log2(m))* siendo m el número de nodos que de como resultado el árbol (dependiendo de los parámetros de terminación). Independiente del número de características de entrada, esto hace de los árboles de decisión un algoritmo rápido.

* Master Thesis
** Facultad De Ciencias Básicas. Escuela de Física.
Maestría en Matemática Aplicada. Director: Raul Ramos Pollán.

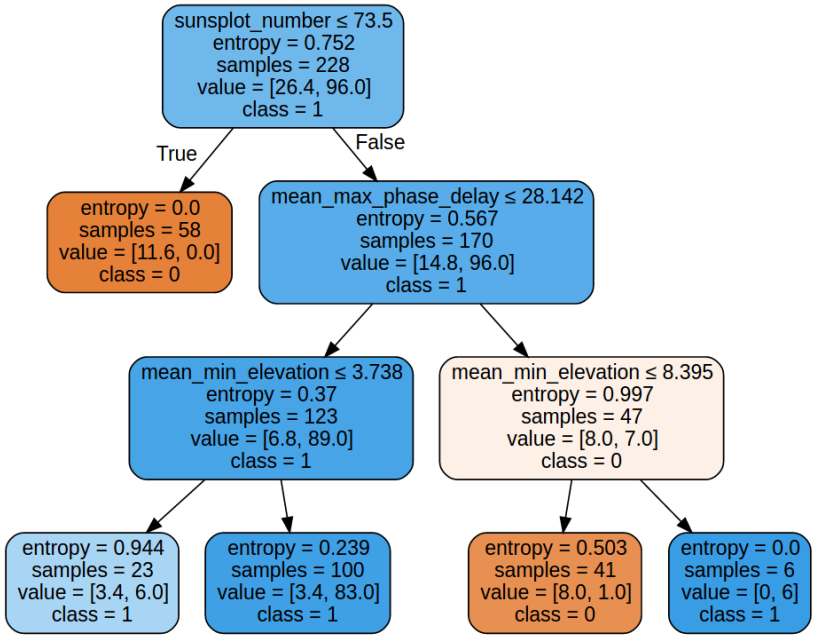

Figura 2: Árbol de Decisión.

Máquinas de vectores de soporte (Clasificación)

Estos algoritmos desarrollados por Vladimir Vapnik y su grupo de trabajo en los laboratorios AT&T en La idea principal es producir un límite de decisión no lineal construyendo un límite de decisión lineal en una versión más grande del espacio de características. Un margen es un hiperplano que separa los datos de diferentes clases, el algoritmo busca el margen que separe de manera óptima los puntos. El mejor margen es aquel con la mayor amplitud posible a cada lado, dentro de todos los posibles hiperplanos separadores. Podemos expresar la regla de decisión en una máquina de soporte vectorial como:

$$\vec{\omega} \cdot \vec{u} + b \geq 0$$

Donde $\vec{\omega}$ es el vector perpendicular al hiperplano separador y $\vec{u}$ es un vector de datos, b es el sesgo. A medida que la proyección de $\vec{u}$ sobre $\vec{\omega}$ se pase de un margen separador, mayor es la probabilidad de que el punto este del otro lado. Para la clasificación binaria denotaremos las etiquetas +1 y -1. $\vec{\omega}$ y b son incógnitas, no tenemos suficientes restricciones para poder calcularlos.

Sean X_1 y X_2 los vectores de soporte (aquellos más cercanos al hiperplano), las restricciones requieren que:

$$\vec{\omega} \cdot \vec{X_1} + b = 1$$
$$\vec{\omega} \cdot \vec{X_2} + b = -1$$

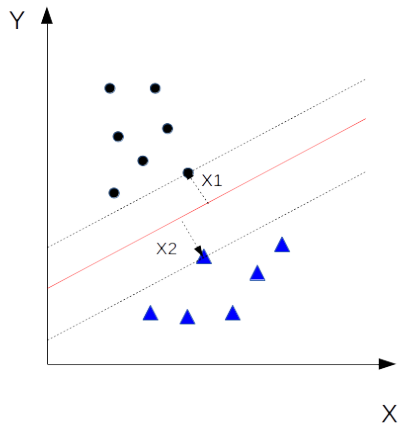

Figura 3: Máquina de Soporte Vectorial para clasificación.

Restando estas dos ecuaciones obtenemos $\vec{\omega} \cdot (\vec{X_1} - \vec{X_2}) = 2$
Si denotamos , X_+, X_- como los ejemplos de la clase +1 y -1 respectivamente tendriamos dos ecuaciones:

$f(u) = \vec{\omega} \cdot \vec{X_-} + b \leq 1$
$f(u) = \vec{\omega} \cdot \vec{X_+} + b \geq 1$

Queremos una función que compacte las dos anteriores en nuestra ecuación de regla de decisión
$y_i\ (\vec{\omega} \cdot \vec{X_i} + b\ ) - 1 \geq 0$, 0 para los puntos pertenecientes a X_-, y ¿1 para los demás.

Dividiendo por la magnitud de $\vec{\omega}$ nos da como resultado la distancia entre las líneas a los lados del hiperplano.

$D = \frac{\vec{\omega} \cdot (\vec{X_1} - \vec{X_2})}{||\vec{\omega}||} = \frac{2}{||\vec{\omega}||}$

Ahora para obtener el hiperplano separador óptimo tenemos un problema de maximización:

$max \frac{2}{||\vec{\omega}||}$, para esto se debe minimizar $\vec{\omega}$. O podemos minimizar la expresión $\frac{1}{2}||\vec{\omega}||^2$ utilizando las restricciones y el método de multiplicadores de Lagrange tenemos:

$L = \frac{1}{2}||\vec{\omega}||^2 - \sum \alpha_i (y_i(\vec{\omega} \cdot \vec{x_i} + b) - 1)$ Donde α_i son los multiplicadores de Lagrange, algunos serán 0. Encontrando las derivadas parciales e igualándolas a cero:

$\frac{\partial L}{\partial \vec{\omega}} = \vec{\omega} - \sum \alpha_i y_i \vec{x_i} = 0$

$\frac{\partial L}{\partial \vec{b}} = \vec{\omega} - \sum \alpha_i y_i \vec{x_i} = 0$

Obtenemos $\vec{\omega} = \sum \alpha_i y_i \vec{x_i}$, reemplazando este valor en la ecuación de Lagrange:

$L = \frac{1}{2}(\sum \alpha_i y_i \vec{x_i}) \cdot (\sum \alpha_i y_i \vec{x_i}) - \sum \alpha_i (y_i((\sum l \alpha_i y_i \vec{x_i})\vec{x_i} + b) - 1)$

$= \sum \alpha_i - \frac{1}{2}(\sum \alpha_i y_i \vec{x_i}) \cdot (\sum \alpha_i y_i \vec{x_i})$

Este resultado nos muestra que el aprendizaje dependerá sólo de los productos internos de pares de muestras. Reemplazando en nuestra regla de decisión tenemos:

$\sum \alpha_i y_i (\vec{x_i} \cdot \vec{u}) + b \geq 0$

Si los datos no son linealmente separables se puede hacer uso del *"kernel trick"* que nos ayuda a sacar provecho de los productos interiores, para mapear los datos a un espacio más conveniente. La maximización dependera de los productos internos.

Sea Φ una transformación en la cual los datos se pueden separar linealmente, queremos maximizar el producto punto en el espacio $\Phi(\vec{x_i}) \cdot \Phi(\vec{x_j})$. El truco consiste en que no necesitamos saber la transformación Φ, Sino la función de kernel K que provee el producto punto de esos dos vectores en en otro espacio.

$K(\vec{x_i}, \vec{x_j}) = \Phi(\vec{x_i}) \cdot \Phi(\vec{x_j})$

Existen diferentes funciones de kernel, entre las más populares se encuentran:

- Lineal: $\langle \vec{x_i} \cdot \vec{x_j} \rangle$
- Polinomial: $(\gamma \langle \vec{x_i} \cdot \vec{x_j} \rangle + r)^d$
- Exponencial: $e^{\gamma ||\vec{x_i} - \vec{x_j}||^2}$
- Sigmoide: $tanh(\gamma \langle \vec{x_i} \cdot \vec{x_j} \rangle + r)$

Redes Neuronales

El término "redes neuronales" comprende un gran número de variados métodos de aprendizaje computacional, estos se pueden aplicar a problemas de clasificación o regresión. Últimamente han vuelto a despertar un gran interés por la comunidad a pesar de que los primeros algortimos datan de los años setenta, debido a nuevos avances en el campo y propiedades que hacen de estos algoritmos soluciones escalables. El nombre proviene del hecho de ser un algoritmo inspirado en cómo se cree es el funcionamiento del cerebro. Así pues cada unidad representará una neurona, y las conexiones entre neuronas (pesos) las sinapsis.

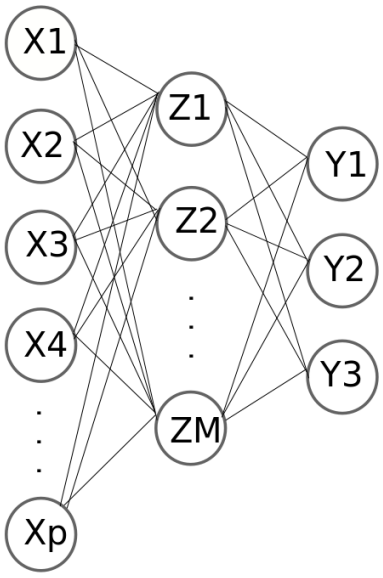

Figura 4: Red Neuronal de una sola capa.

Como toda solución de aprendizaje computacional, lo que se busca es una aproximación. Las redes neuronales crean una función no lineal a partir de combinaciones lineales de entrada. En la figura se aprecia una red sencilla, donde existen tres capas: Capa de entrada con p unidades , una capa oculta y una de salida, con m y K unidades respectivamente. Cuando tratamos un problema de regresión K = 1, para regresión K será el número de unidades con la probabilidad modelada por cada clase k a predecir.

Sea X el conjunto de datos de entrada con p carácterísticas en la capa de entrada, a apartir de estas se derivan nuevas características Z_m a partir de tranformaciones no lineales. Finalmente y_k es modelada como una función de combinaciones lineales de combinaciones de Z_m

$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X)$, m=1,2,...M
$T_k = \beta_{0k} + \beta_k^T Z$ k = 1,2,...,K
$T = (T_1, T_2, ..., T_K)$, $Z = (Z_1, Z_2, ..., Z_K)$
$f_k(X) = g_k(T)$, $k = (1, 2, ..., K)$

La función σ conocida como la función sigmoide, es una función de activación que nos permite pasar la suma de combinciones de pesos y entradas a valores a un rango entre 0 y 1, esta función nos da una medida de que tanto se activa una "neurona".

$\sigma(v) = \frac{1}{1+e^{-}v}$

Las unidades Z_m conforman la capa oculta, este nombre se le dió porque no observamos directamente su comportamiento. Se puede decir que es una expansión

* Master Thesis
** Facultad De Ciencias Básicas. Escuela de Física.
Maestría en Matemática Aplicada. Director: Raul Ramos Pollán.

de la base de las entradas originales. Finalmente f_k serán las probabilidades finales calculadas para cada clase k , a partir de una transformacion no lineal de T.

Los parámetros o pesos α y β deben ser aprendidos a partir de los datos de entrenamiento. El mecanismo clásico para ajustar los parámetros es llamado "propagación hacia atrás".

Ajustando una Red Neuronal

Buscamos los parámetros θ que se ajusten correctamente a los datos para producir la salida (y_k) deseada. Para esto se necesita una función de costo o error que nos cuantifique que tanto difieren las predicciones hechas por la red de la realidad. Así pues para un problema de regresión podemos usar suma del error cuadrado y para clasificación la entropía cruzada.

$R(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_i k - f_k(x_i))^2$ Función de costo de Suma del error cuadrado

$R(\theta) = -\sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} log(f_k(X_i))$ Función de costo de Entropía cruzada.

Nuestro objetivo será encontrar un mínimo en una función de costo. Para este fin se suele usar el método del **Gradiente Descendiente**, gracias a la forma en que se compone el modelo de red.

El gradiente descendiente es un método iterativo por medio del cual se busca el descenso más rapido en una función. A diferencia de otros métodos donde se busca un míimo o máximoglobal, la función de costo puede no ser convexa, y el mínimo local que se encuentre dependerá del valor inicial de los pesos, estos suelen ser inicializados de manera aleatoria.

Para la función suma de error cuadrado se calculan las derivadas parciales respecto a los parámetros:

$\frac{\partial R_i}{\partial \beta_{km}} = -2(y_{ik} - f_k(x_i)) g_k'(\beta_k^T Z_i) Z_{mi}$
$\frac{\partial R_i}{\partial \alpha_{ml}} = -\sum_{k=1}^{K} 2(y_{ik} - f_k(x_i)) g_k'(\beta_k^T Z_i) \beta_{km} \sigma'(\alpha_m^T x_i) x_{il}$

En el gradiente descendiente con cada $r + 1$ iteración se actualizan los parámetros:

$\beta_{km}^{r+1} = \beta_{km}^r - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \beta_{km}^r}$

$\alpha_{ml}^{r+1} = \alpha_{ml}^r - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \alpha_{ml}^r}$, γ : Tasa de aprendizaje.

Podemos reescribir las derivadas parciales:

$\frac{\partial R_i}{\partial \beta_{km}} = \delta_{ki} Z_{ml}$, $\frac{\partial R_i}{\partial \alpha_{ml}} = S_{mi} x_{il}$.

δ_{ki} y S_{mi}son los "errores" del modelo en esa iteración en el resultado y la capa oculta respectivamente. Además note que:

$S_{mi} = \sigma'(\alpha_m^T X_i) \sum_{k=1}^{K} \beta_{km} \delta_{ki}$

Usando las actualizaciones del gradiente descendiente se implementa la famosa propagación hacia atrás que consta de dos pasos:

- Pase hacia adelante: Los pesos son fijados y las predicciones f_k computadas.
- Pase hacia atrás: Los errores δ_{ki} son computados y propagados hacia atrás calculando S_{mi}. Ambos errores son usados para computar actualizar el gradiente descendiente.

Ventajas y Desventajas de las Redes Neuronales

Entre las ventajas de este método se encuentran su simple naturaleza "local", cada capa pasa y recibe información unicamente de las unidades a las que se encuentra conectada, la implementación de las redes en arquitecturas paralelas presentan un gran aumento en la eficiencia en términos de disminución de tiempo de entrenamiento.

El hecho de que la función de costo no es convexa supone una de las desventajas de las redes neuronales. Ya que siempre existe el riesgo de quedarse estancado en un mínimo local no deseado.

3.1.3. Métodos de Ensamble

Un ensamble es un meta-algoritmo, el resultado de unir o agregar un grupo de predictores o aprendices que pueden ser considerados débiles (apenas mejor que un clasificador aleatorio) con el fin de crear un modelo más fuerte. Los métodos de ensamble pueden unir varios clasificadores de igual o diferente tipo. Según el tipo de ensamble se puede disminuir la varianza o el sesgo, dando mejores resultados que el mejor de los clasificadores contenidos, de cierta manera un ensamble convierte un modelo en una característica.

Métodos de Bagging

Su nombre proviene del inglés (bootstrap aggregating), estos métodos usan el mismo algoritmo de entrenamiento para cada predictor o clasificador, pero se entrena en diferentes subconjuntos aleatorios de los datos de entrenamiento, el método de muestreo es con reemplazamiento, de lo contrario sería un ensamble tipo "Pasting". Estos métodos suelen explotar la independencia entre modelos base, permitiendo

reducir el error, disminuyendo la varianza. Para obtener los subconjuntos de entrenamiento se usa bootstrap, para agregar las salidas de los modelos base de hace una votación en el caso clasificación y el promedio en caso de regresión.

Ensamble de Votación

Entrena diferentes algoritmos de clasificación y según la mayoría en el voto o el promedio predice la clase a la cual pertenece un elemento. Estos algoritmos a menudo alcanzan un mejor desempeño que el mejor clasificador en el ensamble.

Bosques Aleatorios

Este algoritmo consiste en entrenar un grupo de árboles de decisión, cada uno sobre un subconjunto aleatorio de los datos de entrenamiento, para hacer predicciones se obtienen las predicciones de cada uno de los decisión y se clasifica en la clase con mayor número de votos.

3.2. Métricas de Desempeño para Modelos de Clasificación

Matriz de Confusión

La matriz de confusión es una tabla que permite visualizar el desempeño de un método de clasificación supervisado, es utilizada en estadística y aprendizaje computacional. Cada fila de la matriz representa las instancias en la clase predecida, y cada columna representa las instancias reales en cada clase.

Cuadro 1: Ejemplo de matriz de confusión

Real/ Resultado	**P**	**N**
P	5	3
N	9	2

A partir de la matriz de confusión se derivan varios términos:

- Condición Positiva (P): Número de positivos en el conjunto de datos
- Condición Negativa (N): Número de negativos en el conjunto de datos
- Verdadero positivo (VP): Número de positivos correctamente identificados.
- Verdadero negativo (VN): Número de negativos correctamente identificados.
- Falso Positivo (FP): Número de negativos identificados como positivos.
- Falso Negativo (FN): Número de positivos identificados como negativos.

Métricas para Clasificación

Podemos definir varias métricas a partir de los conceptos anteriores:

Tasa de verdaderos positivos: La proporción de positivos correctamente identificados como tales.

$$TVP = \frac{VP}{P} = \frac{VP}{VP + FN}$$

Tasa de falsos positivos: La proporción de falsos positivos respecto al número real de negativos.

$$TFP = \frac{FP}{N} = \frac{FP}{VN + FP}$$

Precisión: Proporción de los datos identificados como positivos que realmente pertenecen a la clase positiva.

$$PRE = \frac{VP}{VP + FP}$$

Fβ: Se puede describir como un promedio entre la presición y la tasa de verdaderos positivos, a medida que β aumenta se le da mayor peso a la TVP.

$$F_\beta = \frac{(1 + \beta^2)VP}{(1 + \beta^2)VP + \beta^2 - FN + FP}$$

3.3. Búsqueda de Hiperparámetros

En aprendizaje computacional una de las tareas más comunes es la búsqueda de los hyperparámetros más óptimos para un algoritmo. Los hiperparámetros son el conjunto de parámetros tales como: pesos, tasa de aprendizaje, coeficiente de regularización entre otros valores que según el método de aprendizaje tienen gran impacto en el desempeño de una tarea. Dado un espacio de hyperparámetros se busca la mejor combinación.

Búsqueda en Grid

Manualmente se establece un espacio de búsqueda, sobre este se realiza una búsqueda exhaustiva probando todos los posibles subconjuntos del espacio de hyperparámtros del algoritmo. Un ejemplo de esto sería para el algoritmo de árbol de decisión:

$A \in \{3, 5, 15, 20\}$, $P \in \{gini, entropy\}$

Tendríamos ocho diferentes árboles de decisión con los posibles montajes del algoritmo. Con validación cruzada se entrena cada modelo y aquel que tiene mejor desempeño es elegido como el mejor conjunto de parámetros.

3.4. Sistemas GNSS/ GBAS

Para comprender mejor la necesidad de este proyecto se presenta un resumen del funcionamiento y fuentes de error en los sistemas de posicionamiento satelital y sobre los sistemas GBAS en aviación.

Los sistemas de posicionamiento global (GNSS) hacen posible que un usuario conozca su posición en el globo, en la actualidad los sistemas operando son GPS (Estados Unidos) y GLONASS (Rusia). Otros dos sistemas se encuentran en fase de implementación y despliegue: GALILEO (Union Europea) y Beidu (China). Estos sistemas cuentan con tres segmentos: Espacial, de control y usuario. El segmento espacial consta de satélites que giran alrededor del planeta y envían señales en dirección hacia la tierra, los satélites tienen relojes atómicos de alta precisión, el tiempo que tarda en ser recibida la señal se multiplica por la velocidad de la luz para obtener el pseudorango: distancia estimada entre el satélite y un receptor en la tierra. Las constelaciones de satélites fueron diseñadas para que un usuario en cualquier parte del planeta pueda tener mínimo cuatro satélites en vista en todo momento. Cuando se tienen varios pseudorangos el problema del cálculo de posición se soluciona con una triangulación.

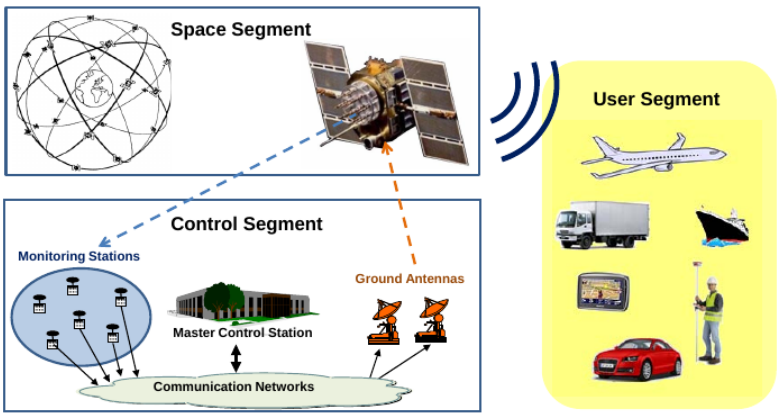

Figura 5: Segmentos GNSS. Fuente: [15]

El segmento de control se encarga de monitorear el funcionamiento del segmento espacial. La fuerza aérea de los Estados Unidos es responsable por la operación de la constelación de GPS. Algunas funciones de este segmento son: corregir errores de las órbitas y relojes de los satélites, además cuenta con múltiples estaciones terrestres distribuidas por todo el planeta para recolectar datos y mandarlos a la estación central que realiza las correcciones de órbitas para una mejor geometría. Por último está el segmento de los usuarios conformado por todos los receptores, este tiene como tarea calcular los pseudorangos y resolver las ecuaciones de posicionamiento. Hay receptores que captan señales en una o más frecuencias, los receptores de doble frecuencia tienen más beneficios a la hora de resolver el retraso ionosférico, pero su precio es mucho mayor. Otros receptores especiales tienen antenas que permiten cancelar el efecto multi-caminos, ruido en la señal, entre otros inconvenientes. Vale la pena mencionar que la gran mayoría de los usuarios hoy en día tienen receptores sencillos de una frecuencia y bajo costo.

3.5. Señales GPS

Las señales GNSS se transmiten por varias frecuencias de la banda L del espectro, la señal está compuesta por:

- Portadora: Es una onda de radio sinusoidal.
- Código de rango: Secuencia de bits que permiten determinar el tiempo de viaje de la señal de radio desde un satélite hasta el receptor.
- Datos de navegación: Mensajes codificados de manera binaria, son datos de efemérides tales como parámetros de órbita que permiten calcular la posición de los satélites, bias de los relojes, correcciones del tiempo, el almanaque, estado de los satélites.

3.6. Observables GNSS

A partir de las señales de sistemas de posicionamiento se obtienen diferentes medidas u observables.

3.6.1. Códigos

El código es modulado sobre la carrier o portadora, permite determinar el tiempo de viaje de la señal de radio desde un satélite hasta el receptor. Existen dos tipos de código: aproximado y preciso.

Código de Adquisición Aproximativa (C/A): Secuencia de uso civil, contiene 1023 bits y se repite cada milisegundo, su longitud de onda es de 293.1 metros. Este código sólo se modula en una frecuencia, y se usa en el servicio estándar de

posicionamiento.

Código de Precisión (P): Código reservado para uso militar y usuarios autorizados, la secuencia se repite cada 266 días, su longitud de onda es de 29.31 metros y se modula sobre L1 y L2. Por lo cual se usa para el servicio de posicionamiento preciso.

Pseudorango de código: Es el rango aparente entre satélite y receptor, que se obtiene a partir del código, este no corresponde a la perfección con la distancia real entre otras razones por la desincronización entre relojes, los retrasos causados por el medio en el que se transporta la señal. El pseudorango de código en la frecuencia f, dado en metros se expresa de la siguiente manera:

$$R_{P_f} = c[t_{rcv}(T_2) - t^{sat}(T_1)]$$

Donde c es la velocidad de la luz, $t_{rcv}(T_2)$ es el tiempo de recepción de la señal medido en la escala del reloj receptor y $t^{sat}(T_1)$ es el tiempo de emisión de la señal medido en la escala del reloj del satélite.

El pseudorango además de la distancia entre receptor y satélite contiene varios errores que afectan su precisión como los errores de relojes, efectos atmosféricos, relativísticos, errores de multicaminos e instrumentales. La ecuación anterior se puede reescribir como:

$$R_{P_f} = \rho + c(\delta t_{rcv} - \delta t^{sat}) + Tr + \alpha_f STEC + K_{P_{f,rcv}} + K_{P_f}^{sat} + M_{P_f} + \varepsilon_{P_f}$$

Donde ρ es el rango geométrico entre los centros de fase de satélite y receptor en tiempos de emisión y recepción ,$c(\delta t_{rcv} - \delta t^{sat})$ es el retraso de los relojes de satélite y receptor, Tr es el retraso causado por la troposfera, $\alpha_f STEC$ es el retraso ionosférico que es dependiente de la frecuencia, $K_{P_{f,rcv}} + K_{P_f}^{sat}$ son retrasos instrumentales que dependen del código y la frecuencia, M_{P_f} es el término del efecto multicaminos y ε_{P_f} representa el ruido de la señal.

3.6.2. Fase Portadora

La señal portadora también nos da una estimación de rango, tenemos el número de ciclos que da la señal antes de alcanzar la tierra. Este se multiplica por su respectiva longitud de onda para obtener distancia en metros. La observación de fase es más precisa que el pseudorango de código, hasta por 2 órdenes de magnitud. Sin embargo, tiene una gran desventaja: ambigüedad, la ambigüedad de fase consiste en un número entero desconocido de longitudes de onda. Adicionalmente este número cambia cada que el receptor pierde la señal, a este fenómeno se le conoce como *cycle slip*, saltos que producen discontinuidades en los rangos. El rango obtenido por

carrier se modela:

$$\Phi_{L_f} = \rho + c(\delta t_{rcv} - \delta t^{sat}) + Tr + \alpha_f STEC + k_{L_{f,rcv}} + k_{L_f}^{sat} + \lambda_{L_f} N_{L_f} + \lambda_{L_f} w + m_{L_f} + \epsilon_{L_f}$$

Esta ecuación es similar al código, adicionalmente se tiene el término de la ambiguedad de ciclos y $\lambda_{L_f} w$ que es el el efecto de la polarización circular de la señal electromagnética.

3.7. Fuentes de Error en GNSS

Las señales enviadas por satélites son afectadas por:

- Retrasos atmosféricos: Retrasos causados por la ionósfera y la tropósfera.
- Error de Multicaminos: La señal puede llegar por diferentes caminos a la antena, esto se da porque tiene objetos reflectores cercanos, en las ciudades este efecto cobra importancia.
- Ruido del receptor: Las medidas de código estan afectadas por ruido, en bajas elevaciones empeora, este se suele suavizar por medio de filtros. La portadora también es afectada por el ruido pero a un orden mucho más bajo. Es común suavizar el código utilizando la fase-portadora. Aunque la señal portadora es más precisa, trae ambiguedades deconocidas.
- Errores de efemérides: El conocimiento de las órbitas y relojes del segmento espacial es de vital importancipara estimar bien la posición, cualquier error afectará la precisión. Esta información es trasmitida en los mensajes de navegación y según el producto usado varía la calidad de la estimación. Estos datos los proveen varias organizaciones, como el Servicio Internacional de GNSS (IGS).
- Efectos de la Relatividad: Debido a la diferencia del potencial gravitacional en la tierra y el espacio los relojes deben ser corregidos.

Error Ionosférico

La ionósfera es la capa de la atmósfera terrestre que está ionizada debido a los rayos X y UV del sol. Se encuentra aproximadamente entre los 80 km y 500 km de altitud, aunque los límites varían según la ubicación geográfica, su concentración de electrones suele ser mayor en la zona ecuatorial. La velocidad de propagación de las señales GNSS en la ionósfera depende directamente de la densidad de electrones. Durante el día por la radiación del sol se liberan electrones, en la noche los electrones

libres se recombinan con iones para producir particulas neutrales, lo cual lleva a una reducción de la densidad de electrones.

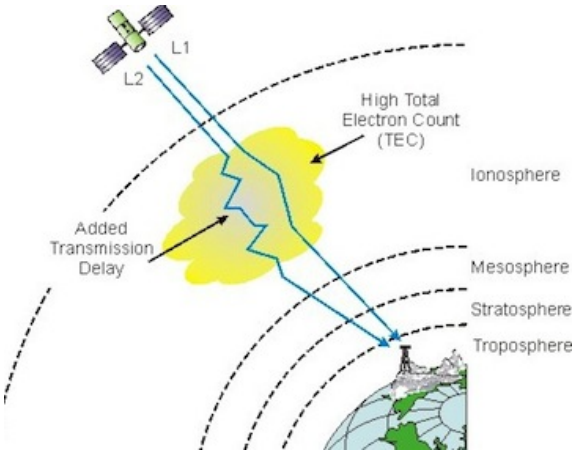

Figura 6: Retrasos causados por la atmósfera. Fuente: `http://xenon.colorado.edu/`

El contenido total de electrones (TEC) y el índice de refracción de este medio dependerá de la ubicación geográfica, la hora del día y la actividad solar. Como la ionósfera es un medio dispersivo la refracción de las señales depende de su frecuencia. En receptores que usan dos frecuencias se puede calcular el retraso fácilmente en condiciones normales, mientras que en los de una frecuencia se necesita aplicar modelos para corregir este efecto, cabe mencionar que actualmente la gran mayoría de los receptores son de una frecuencia.

3.8. Sistemas GBAS

Los sistemas de aumentación terrestres GBAS ayudan a monitorear los sistemas de posicionamiento global (GPS, GLONASS, Galileo, etc.) y proveer correcciones mejorando la calidad del servicio, mejoran la integridad, dan monitoreo en tiempo real y aumentan la precisión por medio de correcciones diferenciales.

Usando una o varias estaciones terrestres de referencia de posición conocida para mejorar el desempeño de los servicios de navegación en area local ya que el SPS (Standard positioning service) por sí solo no cumple con los requerimientos de integridad que necesitan las operaciones de aviación. Se usa para dar soporte en las fases de despegue y aterrizaje de un vuelo.

Al tener un usuario y estación referencia podemos estimar mejor el error de efemérides y el causado por la ionósfera. Ya que el riesgo potencial de estos se correlacionan cuando la distancia entre usuario y estación es corta.

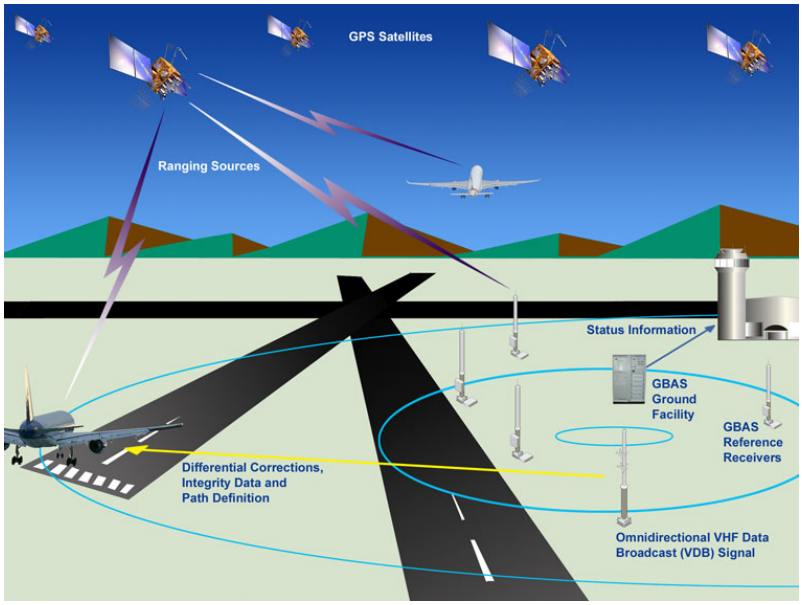

Figura 7: Segmentos GBAS. Fuente: `https://www.faa.gov`

- Terrestre: Monitorea la integridad de las señales satelitales, calcula correcciones diferenciales y provee mensajes con correcciones para el usuario por un canal de alta frecuencia (VDB).
- Espacial: Constelación de satélites (GPS o GLONASS) para uso civil, con opción de usar sistema SBAS si está disponible.
- Segmento aéreo: Usuario con receptor, se encarga de recibir y decodificar las señales GNSS y GBAS, determina la posición, computa la disponibilidad del servicio y desviaciones del camino.

Las correcciones que hace se emiten a traves de una canal de datos de alta frecuencia VDB. La estación puede proporcionar posicionamiento a usuarios que se encuentren a menos de 100 kilómetros de distancia.

4. Estado del Arte

El estudio y modelado de la ionósfera es de vital importancia para que los sistemas GNSS tengan la precisión deseada en todo el globo, a continuación se presenta un resumen de los principales modelos desarrollados para estimar los retrasos causados por este medio, así como también el uso de tecnologías de ineligencia artificial para mejorar las aplicaciones que usan GNSS.

4.1. Métodos de mitigación del error ionósferico

El modelo Klobuchar [7] es aplicado en los receptores actualmente. Desarrolado para el sistema GPS por John A. Klobuchar, reduce el error de rango producido en promedio hasta en un 50 %. Modela la ionósfera como una capa delgada a 350 kilómetros de altura, el retraso de la señal inclinada es calculado del retraso vertical en el punto de perforación de la capa (IPP), esto se hace multiplicando por un factor de oblicuidad. Los coeficientes de Klobuchar y son enviados en el mensaje de navegación para calcular el retraso. Los retrasos verticales estan basados en un valor constante de cinco nanosegundos en las noches, sin embargo en ciertas latitudes se presentan burbujas de plasma durante la noche que este modelo no tiene en cuenta.

El modelo adoptado por el sistema GALILEO para reducir el retraso ionosférico es Nequick, produce un mapa de la ionósfera que depende de variables como la ubicación y el tiempo para estimar la densidad total de electrones (TEC). El algoritmo [4] está disponible para el público, es un modelo basado en datos que ha creado diferentes perfiles de densidad de electrones entre un receptor y un satélite, el valor de TEC se obtiene integrando el perfil. Los coeficientes usados para aplicar la corrección vienen en el mensaje de navegación junto con alertas en caso de que las correcciones no sean confiables.

Como se mencionó anteriormente, para el monitoreo de los aviones civiles los GBAS son la solución más económica y viable a corto plazo, para la vigilancia de amenazas ionosféricas este emplea un “Modelo de amenaza” o un umbral que fue el resultado de un gran estudio ionosférico hecho por [6] a partir de analizar datos de la red de estaciones CORS y WAAS, se seleccionaron algunos días para procesar según el índice de actividad geomagnética registrado , que representa los efectos de partículas solares en los campos magnéticos de la tierra. Los datos de estaciones de doble frecuencia fueron procesados para detectar retrasos ionosféricos y gradientes espaciales entre estaciones con menos de 100 kilómetros de distancia entre sí. Algunos días en condiciones nominales, y otros días donde se esperaba ver gran actividad y tormentas ionósfericas fueron analizados con gran detalle. El proceso es semi-automático, pues los gradientes son validados en la etapa final por un experto que elimina los falsos positivos que puedan quedar al final del proceso. El modelo de amenaza final depende de la elevación y el gradiente calculado, si se supera el umbral la información proveniente del satélite puede estar afectada por un evento ionosférico figura . El resultado de este estudio fue implementado en las estaciones Smart Path de Honeywell para la detección de gradientes en aeropuertos. Sin embargo, se hizo evidente la falta de estrategias para mitigación en latitudes bajas cuando se instaló una estación en Brasil utilizando este modelo. Actualmente el modelo de amenaza se encuentra siendo actualizado para poder cumplir con los requerimientos de aviación en todo el globo según [9].

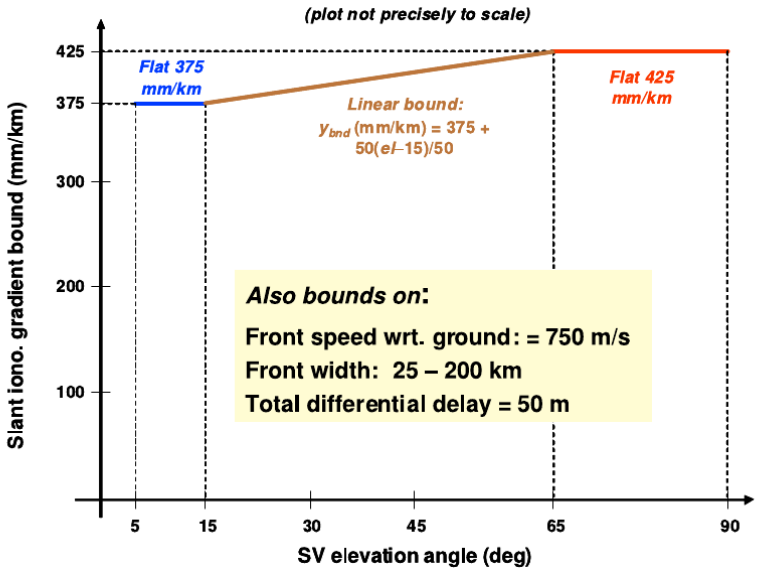

Figura 8: Modelo de Amenaza de CONUS. Fuente: [13]

4.2. Aplicaciones del Aprendizaje Computacional en GNSS

Las técnicas de inteligencia artificial han sido ampliamente aplicadas en diferentes contextos para descubrir patrones y relaciones en todo tipo de datos, GNSS no es la excepción, algunas de sus aplicaciones son: mejorar precisión en entornos donde el efecto multicaminos es la mayor fuente de error, por medio de máquinas de soporte vectorial [12], otros usos más comunes incluyen la predicción de tiempo de viaje usando datos de GPS o recomendaciones de rutas para aplicaciones de tráfico.

El comportamiento de la ionósfera ha sido objeto de diferentes estudios estadísticos como en [17] donde se estudia la compleja relación entre anomalías ionosféricas y eventos de clima espacial, así mismo se caracerizan tipos de anomalías en días sin alteraciones de clima espacial y usa un conjunto de datos de mapas TEC para clasificación binaria usando valores para clusterizar los datos, aprendizaje no supervisado. En otras publicaciones se puede observar el uso de redes neuronales para calcular TEC como en [2] y [11], máquinas de soporte vectorial en series de tiempo para predecir anomalias sismo-ionosféricas a partir de medidas TECU [1]. Además la ionósfera al ser un medio muy estudiado ha propiciado la creación de conjuntos de datos comúnmente utilizados para su inspección por parte de la comunidad científica, un ejemplo de esto es el conjunto de datos donado en 1989 en la web de la Universidad de California en Irvine, que cuenta con 351 instancias y 34 características.

El presente trabajo propone la construcción de un conjunto de datos obtenidos a partir de descriptores de señales de receptores doble frecuencia junto a datos de clima espacial, en diferentes periodos de tiempo, de esta manera se quiere estudiar

la posibilidad de detectar una anomalia teniendo en cuenta muchas variables como la hora del día, la actividad solar sin importar que esta sea alta o baja y las peculiaridades de los retrasos ionosféricos. La aplicación concreta del método propuesto es reducir el número de falsas alarmas que disminuyan la continuidad del servicio GNSS para aviones civiles.

5. Metodología

Para cumplir con los objetivos del proyecto se planteó el uso de las metodologías ágiles, estas metodologías son flexibles y permiten realizar ajustes durante la realización de un proyecto. Cada etapa o paquete tiene sus objetivos junto a una lista de actividades a realizar.

Paquete 1: Revisión del Estado del arte

Actividades

- **A1.1:** Revisión de estado del arte sobre métodos de mitigación del error ionosférico.
- **A1.2:** Revisión de estado del arte del uso de métodos de aprendizaje computacional en GNSS.

Paquete 2: Métodos semi-automáticos

Actividades

- **A2.1:** Definición e implementación del cálculo preciso de retrasos ionosféricos por estación GNSS.
- **A2.2:** Definición e implementación del cálculo de los gradientes ionosféricos entre pares de estaciones.
- **A2.3:** Procesado masivo de datos GNSS red de estaciones seleccionadas.
- **A2.4:** Validación manual de los gradientes no conformes con el modelo de amenazas ionosféricas.

Paquete 3: Desarrollo de Métodos de aprendizaje automático para la solución del problema.

Actividades

- **A3.1:** Creación de datasets de gradientes ionosféricos para aprendizaje supervisado.
- **A3.2:** Diseño de descriptores para retrasos ionosféricos y gradientes.

---

* Master Thesis
** Facultad De Ciencias Básicas. Escuela de Física.
Maestría en Matemática Aplicada. Director: Raul Ramos Pollán.

- **A3.3:** Definición de métodos de preprocesado y clasificación.

Paquete 4: Evaluación de desempeño e implementabilidad de los métodos desarrollados.

Actividades

- **A4.1:** Evaluación experimental del desempeño.
- **A4.2:** Estimación de costes computacionales para sistemas en operación.
- **A4.3:** Evaluación de la integrabilidad de los métodos desarrollados en redes de monitoreo ionosférico.

6. Desarrollo

La aparición de nuevas redes de receptores GNSS de doble frecuencia en diferentes partes del globo permitió realizar una recolección masiva de datos satelitales para nuestro estudio. Un requerimiento para el el cálculo de anomalías detectadas en GBAS es que existan diversos pares de estaciones a menos de 100 kilómetros de separación, ya que a mayor distancia las medidas de gradientes ionsoféricos se deterioran (si están muy separados no se observa la misma sección de ionósfera) y no son medidas confiables. Los productos necesarios para calcular señales de retrasos ionosféricos son archivos RINEX que fueron extraídos de tres redes en diferentes ubicaciones.

6.1. Conjunto de Datos

- **Conterminous United States (CONUS):** Los datos de esta red fueron usados para implementar el modelo de amenaza conus, además de que fueron clave para el proceso de validación de la herramienta desarrollada para computar gradientes, explicada en la siguiente sección. Los datos de esta red corresponden a la región de norte america y Alaska, latitudes medias y altas en donde las tormentas ionosféricas son menos fuertes. Estos datos se pueden acceder en `https://www.ngs.noaa.gov/CORS/coords_alt.shtml`.

- **Red GNSS de Monitoreo Continuo del Ecuador (REGME):** Cuenta con 45 estaciones de sensores multiconstelación en la región del ecuador y las islas galápagos, el Instituto Geográfico Militar junto con otras entidades públicas mantienen la red, sus datos son de fácil acceso para investigaciones y aplicaciones científicas. El área que cubre la red es de especial interés pues la región ecuatorial tiene la mayor actividad ionosférica por estar más cerca del sol, lo cual afecta la precisión de las aplicaciones de GNSS en latinoamérica donde no se cuenta aún con SBAS.

- **Red Andaluza de Posicionamiento (RAP):** Esta red de estaciones en la region de Andalucía en España, provee correcciones diferenciales y cuenta con 22 estaciones.

- **Productos del Servicio Internacional de GNSS (IGS):** Además de los datos de receptores GNSS de diferentes redes en el planeta, provee una variedad de productos que fueron extraidos del sitio web de IGS. Efemérides presentes en los mensajes de navegación y datos para estimación del ruido en satélites según la disponibilidad en cada momento del procesamiento de los gradientes ionosféricos.

- **Indices de actividad solar y del campo magnético de la tierra.**

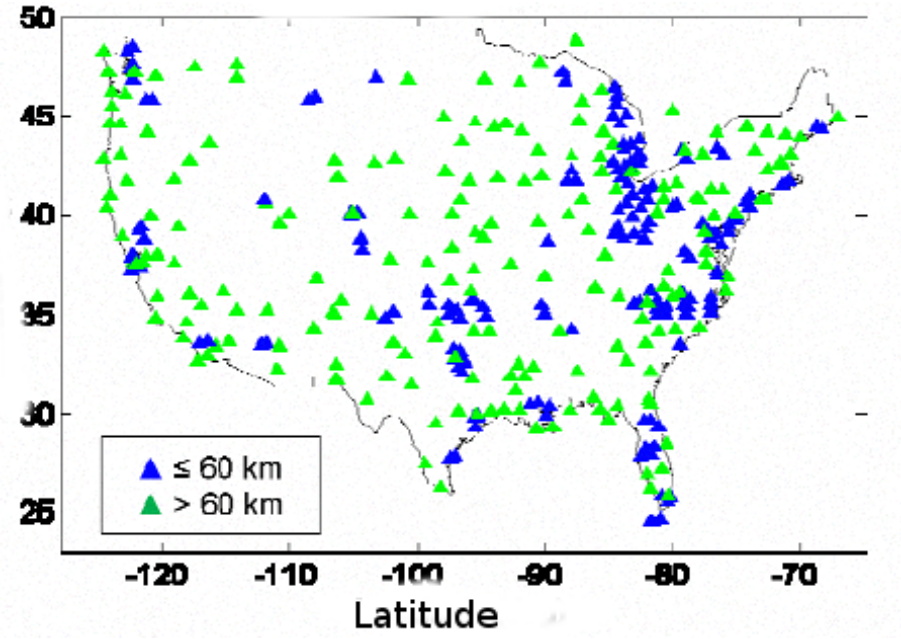

Figura 9: Red de estaciones CONUS. Fuente: [8]

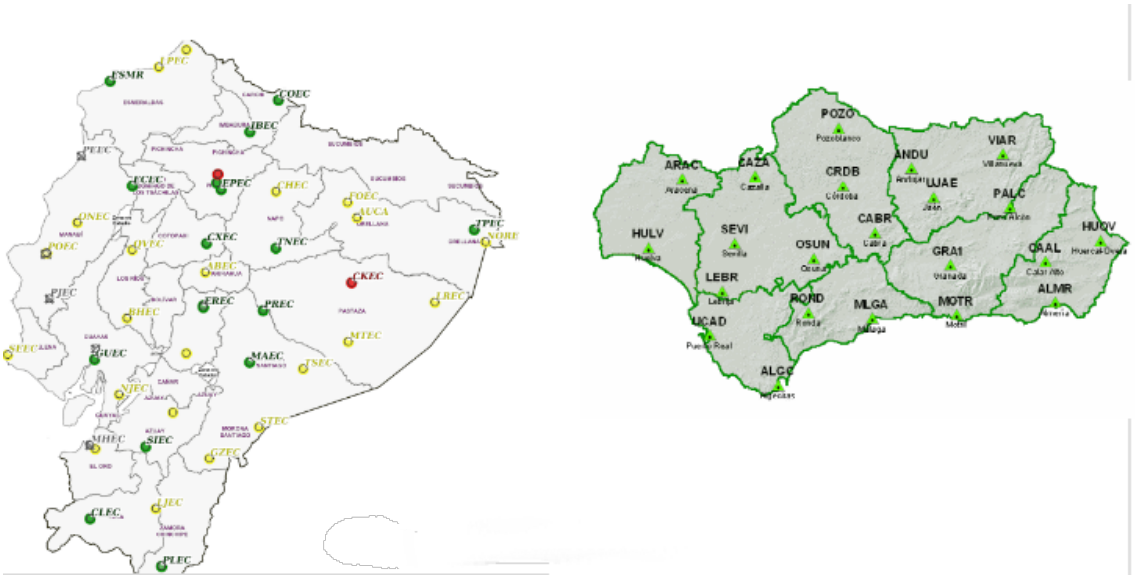

Figura 10: Red del Ecuador REGME (izquierda). Red Andaluza de posicionamiento (derecha).

6.2. Metodología de Cálculo de Gradientes Ionosféricos

El objetivo de la metodogía desarrollada por el Instituto Avanzado de Ciencia y Tecnología de Corea junto a la Universidad de Stanford [6] es calcular retrasos ionosféricos precisos a partir de estas dos señales. Durante esta etapa del proyecto la

metodología fue implementada y validada para luego ser usada sobre una gran cantidad de datos para la conformación de un conjunto de datos. El retraso ionosférico inclinado en la señal GNSS puede ser calculado en receptores doble frecuencia tanto en la señal portadora como en el código, ya que es el mismo en ambas frecuencias de modo que es una ecuación con una sola incógnita.

$$I_\rho = \frac{\rho_{L2}-\rho_{L1}}{\gamma-1} = I + \frac{c}{\gamma-1}(IFB_i + \tau^k_{gd}) + \varepsilon_\rho$$
$$I_\phi = \frac{\phi_{L1}-\phi_{L2}}{\gamma-1} = -I + \frac{c}{\gamma-1}(IFB_i + \tau^k_{gd}) + \frac{N_{L2}-N_{L1}}{\gamma-1} + \varepsilon_\phi$$
$$\gamma = \frac{f^2_{L1}}{f^2_{L1}}$$

Donde c es la velocidad de la luz en el vacío, IFB_i y τ^k_{gd} son los sesgos interfrecuencias para el receptor y el satélite k. El retraso ionosférico derivado de las medidas de código es mucho más ruidoso que el generado a partir de la fase $\varepsilon_\rho >> \varepsilon_\phi$. También se observa en la ecuación de las ambiguedades enteras N_1 y N_2 causadas por la ionósfera en cada frecuencia.

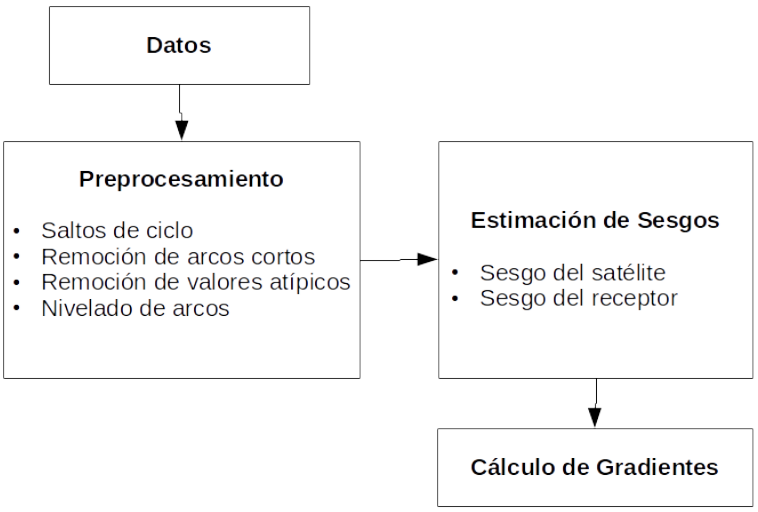

Figura 11: Cálculo de gradientes ionosféricos.

Preprocesamiento

Después de calcular estas señales se aplican varios pasos para obtener una señal de fase final sin ruido ni ambiguedades. Se detectan saltos de ciclo en la fase, un salto de ciclo se da cuando se pierde la cuenta del número de ciclos de la señal, se observa como una interrupción. La presencia de un salto de ciclo se determina si dos puntos adyacentes estan separados menos de una hora en el tiempo y su diferencia supera un umbral (0.8 ó 2.5 metros dependiendo de la actividad solar en el día). También se tiene en cuenta un índice de pérdida de señal (LLI) presente en el mensaje de observación RINEX, finalmente la ausencia de observables de código o fase en cualquiera

de las dos frecuencias L1/L2 es también considerada un salto de ciclo. Cada salto es usado para segmentar la señal en "arcos", cada uno de los cuales será procesado independientemente.

Los valores atípicos en la señal son comunes, deben ser eliminados para lo cual se emplean dos métodos: Un ajuste polinomial P_{fit} y el cálculo de un índice de pertenencia o *outlier factor* OF. En el primer método se calculan los residuales de los puntos del arco menos un un polinomio de grado dos ajustado a la curva. Se computan los residuales diferenciales, si la diferencia más grande entre dos puntos adyacentes supera los 0.8 metros, el punto es clasificado como un valor atípico potencial. El outlier factor es computado al tiempo para cada punto p en el tiempo t_p del arco.

$$OF(t_p) = \sum_{q \in Adyacentes} w_{pq}.|I_p - I_q|$$

$$w_{pq} = \frac{1/|t_p - t_q|}{\sum_{r \in Adyacentes} 1/|t_p - t_r|}$$

Donde los puntos adyacentes son todos aquellos centrados alrededor de cinco minutos del punto p. Si los dos métodos coinciden el punto es eliminado, este proceso se repite hasta que no quedan más valores atípicos.

La señal de retraso ionosférico derivada del código I_ϕ se filtra y a continuación, para cada arco se calcula un factor de nivelación L para la fase. De esta manera se obtienen señales de fase sin ambiguedades, haciendo una suma ponderada de las diferencias entre los retrasos usando la elevación del satélite, para que se le de mas peso a las observaciones con un ángulo mayor, ya que a bajas elevaciones se afecta la calidad de las señales.

$$L = \frac{\sum_i^N (I_\rho(t_i) - I_\phi(t_i)) \sin^2 el_i}{\sum_i^N \sin^2 el_i}$$

$$I_{\phi nivelado} = I_\phi + L = I + \frac{c}{\gamma - 1}(IFB + \tau_{gd})$$

Estimación de Sesgos

En este punto, aún existen errores en el retraso, los sesgos interfrecuencia mencionados anteriormente. El sesgo del satélite es calculado a partir de un producto disponible en los FTP del IGS. Asumiendo que estos son correctos se procede a calcular el sesgo del receptor a partir de método de [10].

Cálculo de Gradientes

Finalmente, los gradientes ionosféricos entre pares de estaciones son calculados simplemente restando los retrasos ionosféricos observados por cada receptor en una

* Master Thesis
** Facultad De Ciencias Básicas. Escuela de Física.
Maestría en Matemática Aplicada. Director: Raul Ramos Pollán.

determinada época, dividido entre la distancia que los separa.

$\nabla I(t) = \frac{|I_i^k(t) - I_j^k(t)|}{||x_i - x_j||}$

Después de este proceso queda un gran número de candidatos gradientes que no son realmente causados por una anomalía en la ionósfera sino por otras causas. Se pone atención a aquellos mayores a 300 mm/Km, para tratar de filtrar los falsos positivos proponen varios mecanismos para descartarlos:

- **Retraso negativo:** Si alguna de las estaciones muestra retraso menor que cero.
- **Sesgo excesivo:** Un gradiente está sesgado si no varía en el tiempo, es señal de que un receptor está defectuoso. Se calcula la media de los puntos del gradiente en el subarco donde se encuentra el candidato, si la diferencia entre el punto máximo y la media es menos de 50 mm/Km el gradiente es descartado.
- **Comparación con Gradiente L1:** Se calcula el gradiente en una sola frecuencia (L1) con el código y fase para comparar con el gradiente doble frecuencia, si hay gran discrepancia entonces las medidas tienen fallas. Los puntos alrededor de 1.5 horas del pico del gradiente son seleccionados. Si el número de puntos para los cuales la diferencia excede los 150 mm/Km es mayor que cinco entonces el candidato es eliminado.

Por último se hace una validación manual para los gradientes restantes que consiste en graficar los que han pasado la validación de filtros para comprobar su validez.

Los datos usados para crear el modelo de amenaza son de la red CONUS, según este estudio la validación manual es necesaria ya que existen casos muy diferentes entre sí [8], esta metodología probada con datos de diferentes latitudes exhibe muchos más falsos positivos por lo cual en las siguientes secciones se explican los pasos para obtener un algoritmo que filtre efectivamente los falsos positivos restantes. En el estudio [14] se usa la misma metodología de cálculo de gradientes con algunas modificaciones, además se cuentan aquellos gradientes mayores de 50 mm/Km en la red REGME del Ecuador en los años 2013 y 2014, sin filtrar los días por indices de clima espacial reportados por el NOAA ya que se quiere evaluar gradientes ionosféricos en los días que no muestran indices con alto de grado actividad geomagnética. Los resultados confirman la presencia de gradientes verdaderos en horas de la noche y madrugada producidos por burbujas ionósfericas, que ocurren muy amenudo en las latitudes bajas.

6.3. Extracción de descriptores y caracterización de eventos ionosféricos

El conjunto de datos está conformado por 305 registros llamados "eventos". Un evento (ver figura 12) es una agrupación de gradientes que han pasado el proceso

automático de validación, cada evento tiene 45 variables o características descriptoras. Los gradientes que forman un evento ocurren en una ventana de tiempo de 30 minutos y con el mismo satélite, un evento debe estar conformado de mínimo dos pares de gradientes. De esta manera es más robusta la hipótesis de que las señales son causadas por una anomalía ionosférica y no por otras causas.

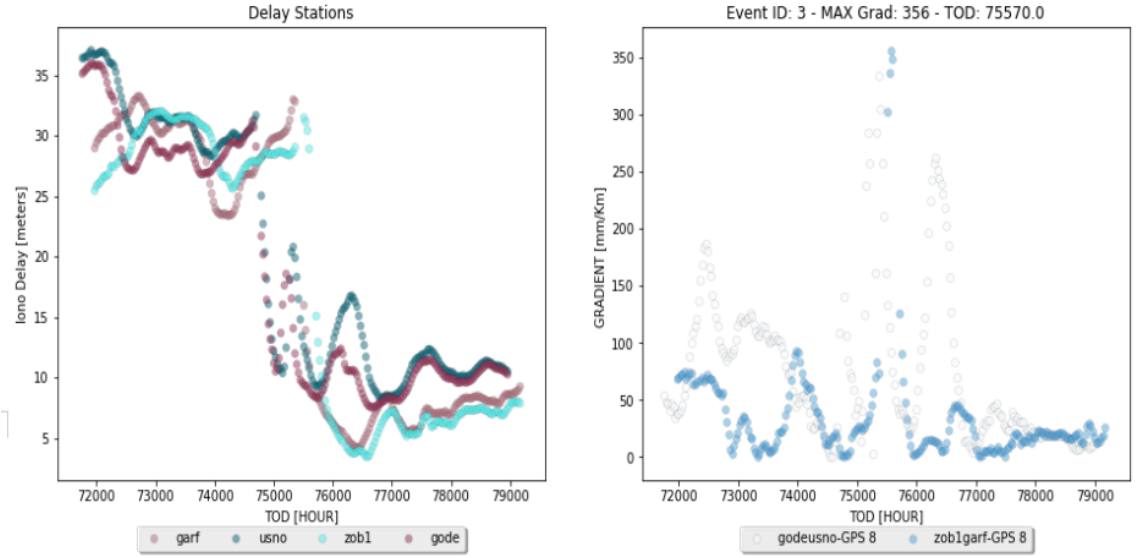

Figura 12: Señales de un evento ionosférico. A la izquierda: Retrasos ionosféricos en la red conus presenciados por 4 estaciones. Derecha: El evento ionosférico en la red CONUS, más de dos gradientes validan la presencia de un evento.

La mayoría de las características son descriptores de las señales presentes en un evento como el de la figura 12 . Similar a [14], no se usan los índices geomagnéticos para filtrar los días de más actividad de clima espacial, tan solo estos índices pasan a ser otra característica descriptiva de nuestro conjunto de datos.

Variables predictoras

Metadatos: Día del año, Año, ID identificador del Evento.

Relacionadas con el retraso ionosférico en estaciones doble frecuencia:

- Tiempo (Hora del día)
- Número de estaciones, número de pares.
- Promedio de la elevacion del satélite, menor elevación, mayor elevación.
- Señal de retraso en fase: Promedio, desviación estándar, valor máximo, valor mínimo, curtosis.

- Señal de retraso en código: Promedio, error cuadrado medio entre el código y la fase.
- Señal de gradiente: Promedio, valor máximo, valor mínimo, curtosis.

Relacionadas con datos de clima espacial:

- Indice Kp: Es un índice usado para caracterizar la magnitud de las tomentas geomagnéticas. Un índice mayor de 4 implica una perturbación en el campo magnético de la tierra. Se reporta cada tres horas por el NOAA.
- Indice A: Indice de actividad geomagnética observada en un día. Un número mayor que 30 indica perturbaciones locales.
- Número de manchas solares: Aparecen como puntos oscuros en la fotosfera. Son regiones de menor temperatura causadas por concentraciones del flujo geomagnético.
- Área de manchas solares.
- Radio flux Penticton 10.7 cm: Valor preliminar observado en Penticton, Canadá. Los valores están en unidades de flujo solar $10^{-}22W/m^2/Hz$.
- Erupciones solares: Incluyen un gran espectro de emisiones o liberación de energía que pueden alcanzar la tierra uno o dos días después de expulsadas. Rayos-x, erupciones S y C.

Además de las variables de entrada originales se agregaron variables en un proceso de *feature engineering*: Variables discretizadas a partir de las variables originales, variables cruzadas, por ejemplo *gradiente alto - elevación de satélite muy baja*, al concatenar estas variables podemos obtener datos con mayor representabilidad de cada evento, lo que permite mejorar el desempeño de los algoritmos.

- Gradiente: Bajo, Medio, Alto
- Retraso: Bajo, Medio, Alto
- Elevación: Muy baja, media, alta.
- Elevacion-gradiente: Concatenacion de los niveles gradiente y elevacion.

El valor de cada característica fue probado mediante varios métodos: Análisis de la varianza fue uno de ellos. En la figura se aprecia la correlación de las variables de entrada con la variable a predecir.

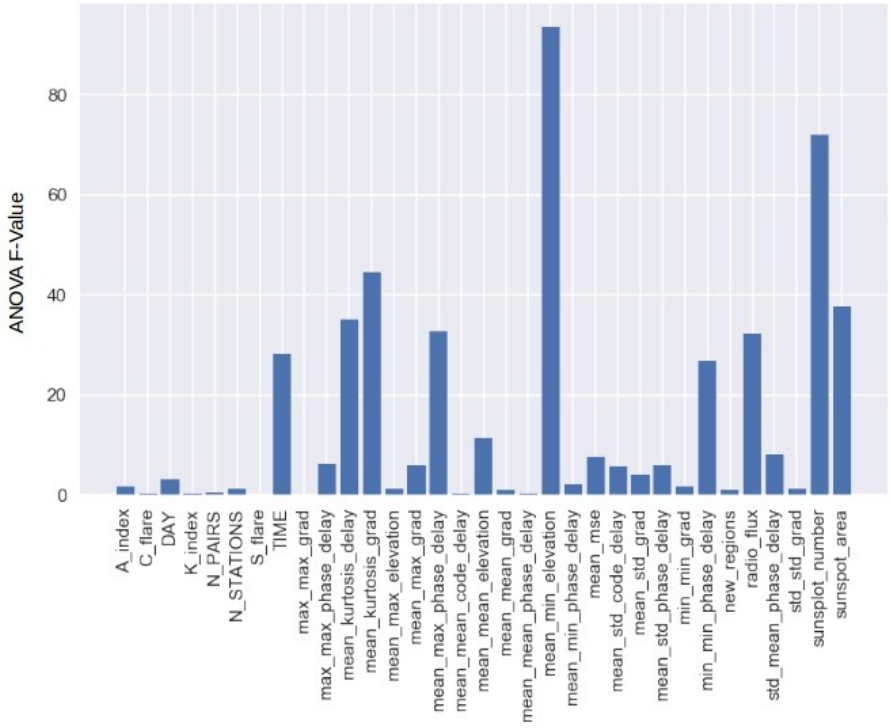

Figura 13: Correlación de las características de entrada.

Se aprecia que la elevación esta fuertemente correlacionada con los gradientes extremos como se esperaba. El número de manchas solares, la kurtosis del retraso ionosférico y la hora del día estan entre las variables que más podrían aportar a un modelo. Contrario a lo establecido en investigaciones pasadas los índices de actividad K y A no parecen muy relacionados con grandes gradientes ionosféricos.

Etiqueta

De acuerdo con el objetivo de clasificación tenemos una etiqueta binaria que indica si el evento ionosférico es verdadero (1) o falso (0). Esta etiqueta fue asignada a mano realizando una inspección de la gráfica de cada evento, tal como se hace en la metodología de validación original.

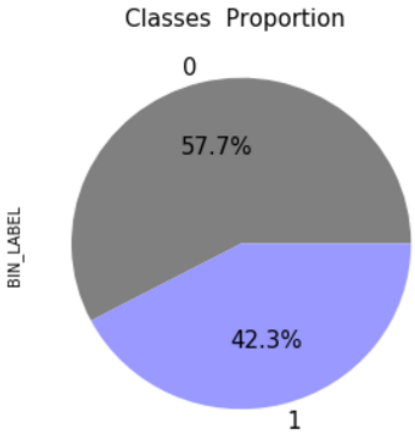

Figura 14: Clases en el conjunto de datos. 0: Negativo, 1: Positivo.

En la práctica son pocos los casos de verdaderos positivos comparados con los negativos, la anterior figura es el resultado de descartar elementos de la clase 0 para poder tener un conjunto de entrenamiento más balanceado ya que el número de instancias de cada clase puede afectar negativamente el rendimiento de muchos algoritmos.

6.4. Montaje experimental: Desarrollo y evaluación de Métodos de Aprendizaje Computacional para Validación de eventos

Después de la conformación del conjunto de datos con las características descritas en la anterior sección se configuró un flujo de trabajo para la construcción del mejor modelo de clasificación binaria. El conjunto original de 305 elementos fue dividido en dos: Conjunto de entrenamiento (75 %) y conjunto de prueba (25 %), por medio de muestreo estratificado conservando la proporción de las clases. El conjunto de entrenamiento es usado para ajustar un algoritmo probando todas las combinaciones de una grilla de parámetros, usando validación cruzada con cuatro folds. Es aquí donde se hace la selección del mejor modelo en entrenamiento, se entrena usando por completo el conjunto de entrenamiento y se prueba en los ejemplos dejados por fuera. Por la aplicación del clasificador en contextos GBAS la métrica elegida para evaluar el desempeño de los algoritmos es la métrica f2, pues se requiere de alta presición y por ende bajo número de falsos positivos al mismo tiempo garantizando alta tasa de verdaderos positivos. El proceso se repite cien veces (validacion cruzada, entrenamiento y seleccion del mejor modelo) para de esta manera garantizar una robustez estadística, y que el modelo final sea el mejor de todas las posibles configuraciones.

En el caso de algunos modelos como máquinas de soporte vectoriales y el perceptrón multicapa el escalamiento ó la normalización de los datos es conveniente para que

las variables no tengan diferentes rangos y a su vez aquellas de mayor magnitud no predominen sobre el resto. Además de ayudar a que la resolución de los parámetros pueda converger más rápido. En las siguientes tablas se aprecia para cada algoritmo el espacio de parámetros usado y la mejor configuración junto al porcentaje en la métrica de desempeño f2.

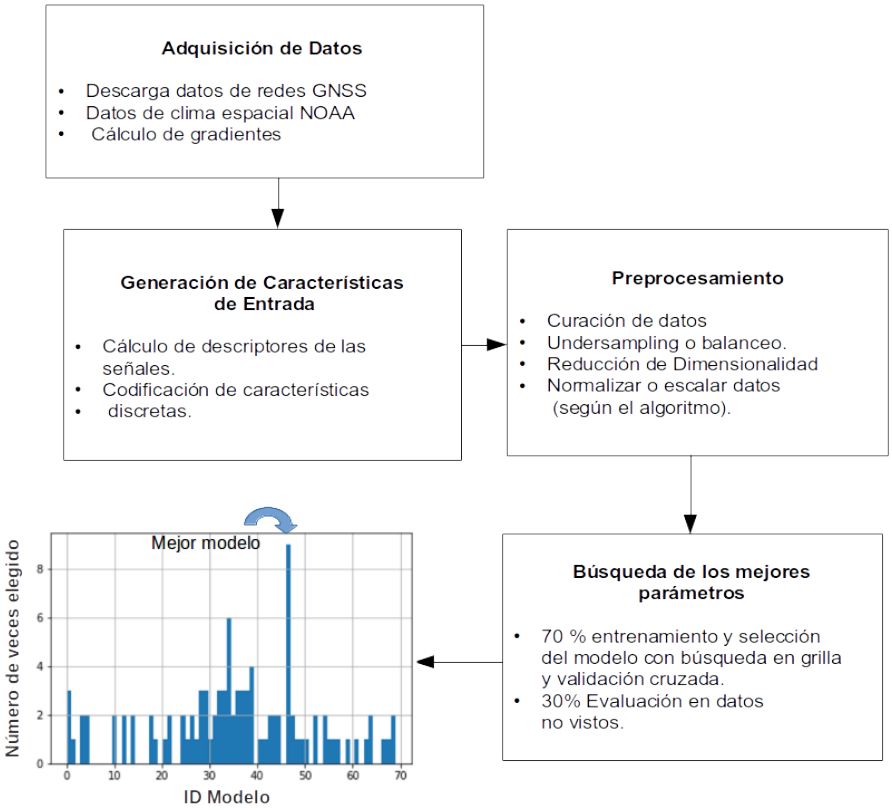

Figura 15: Flujo de trabajo de aprendizaje supervisado.

Modelos seleccionados

Algoritmo	Parámetros	Modelo seleccionado	F2 test
Árboles de Decisión	Criterio: {Gini, Entropía}, Ponderación de las clases:{[0.5,1], [0.2,1], Ninguno} Profundidad: {0, 3, 6, ... 43}	Criterio: Entropía, Ponderación de las clases: [0.2,1] Profundidad:,3	89,05 %
Regresión Logística	Solucionador: {newton-cg, lbfgs, sag, liblinear} Regularización C: {0.1,0.2,...,1}	Solucionador: newton-cg C= 0.10.868107	86,81 %
Clasificador Bayesiano Ingenuo	No necesita.	No necesita.	83,059 %
Análisis de discriminate lineal	Solucionador: lsqr, svd	Solucionador: lsqr	88,78 %
Máquinas de Vectores de soporte	Kernel:{exponencial, sigmoide, lineal} Regularización: { 0.5,0.625,...,1}	Kernel: exponencial Regularización:1	92,83 %
Perceptrón multicapa	Número de capas ocultas:{,5, 25, 45, ...,105} Regularización: {0.001,0.01,1x10-5}	Número de capas ocultas: 105 Regularización: 0.01	88,44 %

Cuadro 2: Resultados de Modelos de aprendizaje.

Algoritmo	Parámetros	Modelo seleccionado	F2 test
Ensamble heterogéneo	Regularización C: {1.0, 0.5, 0.1} Capas ocultas:{5,10,20,40} Máxima profundidad (árbol):{3,10,20} Anova k: {5,10,20,30}	C: 0.1 Capas ocultas: 40 Máxima profundidad: 3 Anova k: 30	87,354 %
Bagging Bayesiano Ingenuo	Máximo porcentaje de características: {0.3,0.5,1} Número de estimadores:{5,10,15}	Máximo porcentaje de características: 0.5 Estimadores:,15	85,457 %
Bagging Regresión Logística	Máximo porcentaje de características: {0.3,0.5,1} Número de estimadores: {5,10,15} Regularización C: {1.,0.5,0.2,0.1}	Máximo porcentaje de características: 0.5 Estimadores: 15 C:1.0	85,857 %
Bagging SVC	Máximo porcentaje de características: {0.3,0.5,1} Número de estimadores: {5,10,15} Regularización C:{1.,0.8,0.5,0.2,0.1}	Máximo porcentaje de características: 0.5 Estimadores: 15 Regularización C: 1	92,081 %
Bosques aleatorios	Criterio: {Gini, Entropía}Número de estimadores:{5,10,15,20} Ponderación de las clases:[{0: 0.5, 1: 1.0}, balanceados] Máxima profundidad:{3,5,10,20}	Criterio: Entropía Número de estimadores: 20 Ponderación de las clases: {0: 0.5, 1: 1.0} Máxima profundidad:3	93,380 %

Cuadro 3: Resultados de Ensambles

Curvas de Aprendizaje

Las curvas de aprendizaje son una representación gráfica que muestra la evolución de un algoritmo de aprendizaje computacional a medida que aumenta la experiencia (datos). Normalmente se grafican dos líneas que corresponden al porcentaje de acierto en los datos de entrenamiento y al porcentaje en datos no vistos. Los algoritmos tienen mejor desempeño en los datos usados para el entrenamiento, sin embargo cuando la brecha entre entrenamiento y test es grande se diagnostica el modelo con sobreajuste. De lo contrario, si tanto test como entrenamiento se encuentran por debajo de la expectativa de clasificación entonces el modelo no esta bien ajustado y se necesita mayor complejidad o más características para mejorar la clasificación. A continuación se presentan las curvas de aprendizaje que muestran los resultados de aquellos modelos seleccionados como los mejores con la metodología de trabajo.

Árboles de Decisión

Figura 16: Curva de aprendizaje para el modelo Árbol de decisión con la métrica f2.

Regresión Logística

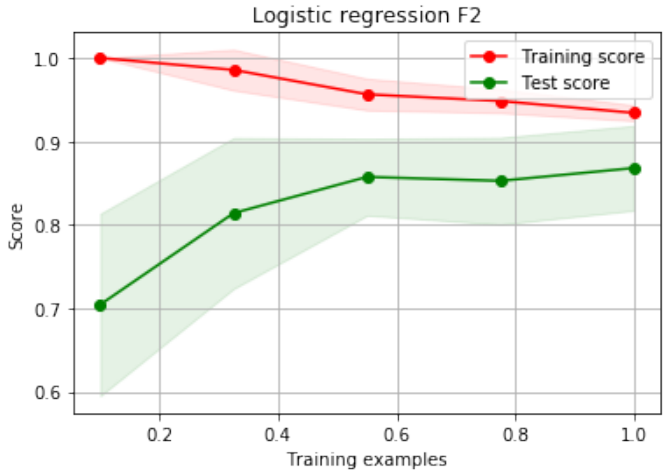

Figura 17: Curva de aprendizaje para el modelo Regresión Logística con la métrica f2.

Clasificador Bayesiano Ingenuo

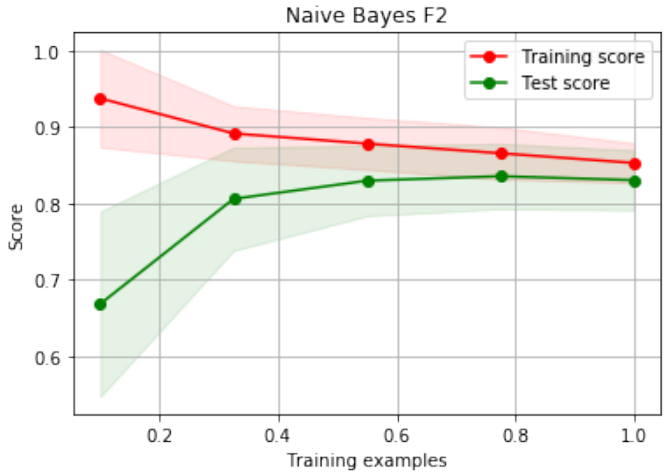

Figura 18: Curva de aprendizaje para el modelo Clasificador Bayesiano Ingenuo con la métrica f2.

Análisis de Discriminante lineal

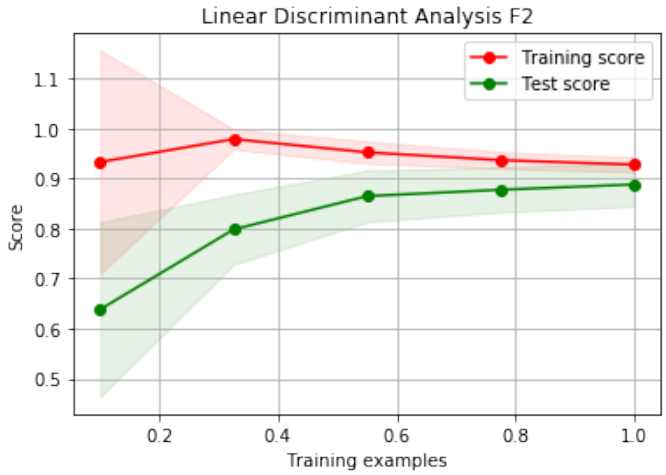

Figura 19: Curva de aprendizaje para el modelo Análisis de Discriminante lineal con la métrica f2.

Máquinas de Vectores de Soporte

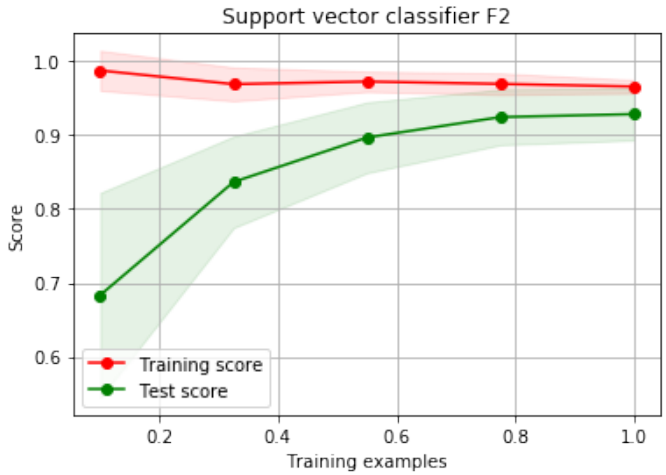

Figura 20: Curva de aprendizaje para el modelo Máquina de soporte vectorial con la métrica f2.

Perceptrón Multi-capa

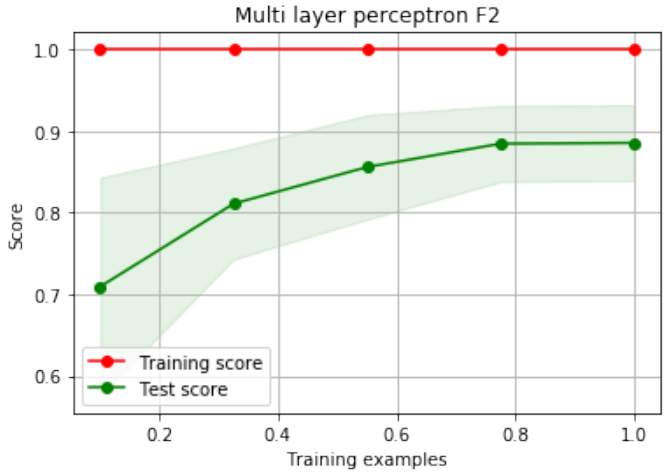

Figura 21: Curva de aprendizaje para el modelo Perceptrón multicapa con la métrica f2.

El modelo de clasificador bayesiano ingenuo (ver figura 18) muestra un ajuste similar en entrenamiento y prueba, mientras que la brecha en el modelo de perceptrón multicapa (figura 21) es claramente un caso de sobreajuste. Los modelos seleccionados muestran resultados satisfactorios, a medida que aumenta la experienia la métrica se mantiene entre el 85 % y el 90 %.

Ensamble Heterogéneo

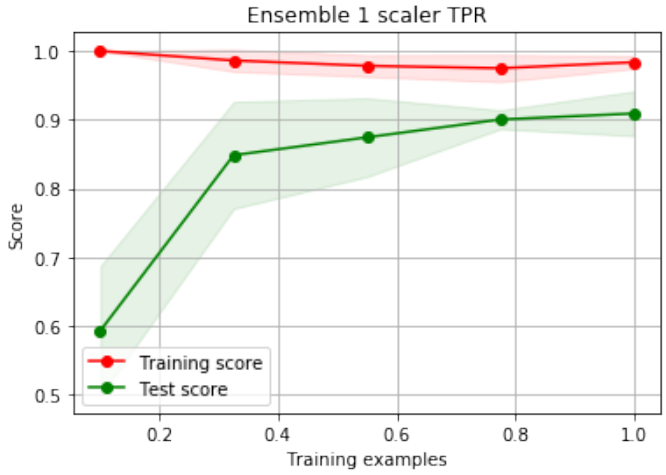

Figura 22: Curva de aprendizaje para el modelo de Ensamble heterogéneo con la métrica f2.

Bagging Bayesiano Ingenuo

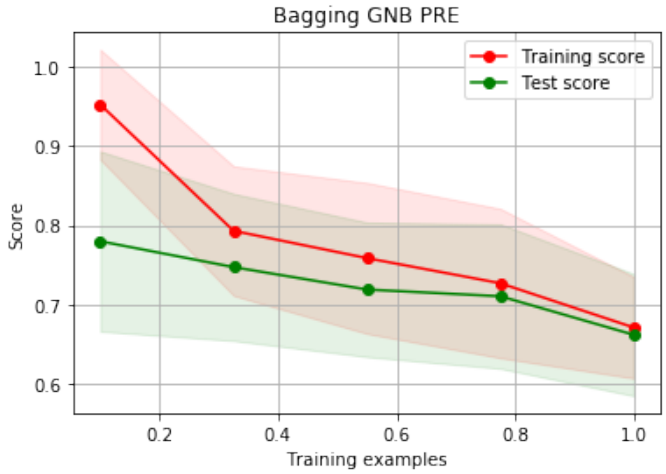

Figura 23: Curva de aprendizaje para el modelo Bagging Bayesiano Ingenuo con la métrica f2.

Bagging Regresión Logística

Figura 24: Curva de aprendizaje para el modelo Bagging Regresión Logística.

Bagging Máquinas de Vectores de Soporte

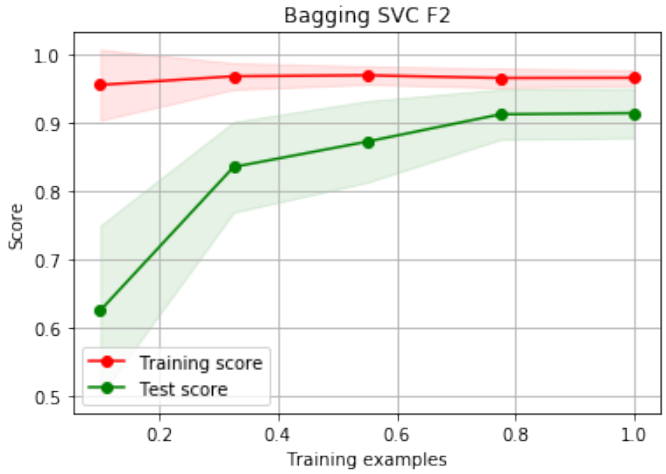

Figura 25: Curva de aprendizaje para el modelo Bagging Máquinas de Vectores de Soporte con la métrica f2.

Bosques Aleatorios

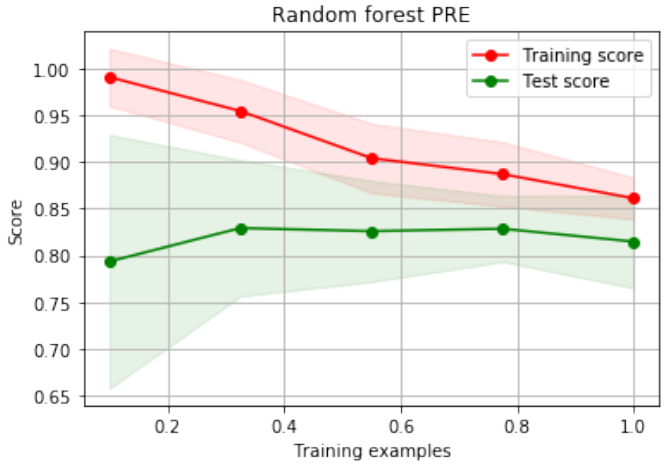

Figura 26: Curva de aprendizaje para el modelo Bosques Aleatorias con la métrica f2.

Los ensambles de un solo tipo de clasificador muestran diferentes comportamientos pero ninguno supera nuestro ensamble heterogéneo, el cual tiene excelentes resultados tanto en entrenamiento como en test. Incluso se puede apreciar que el algoritmo de Bagging usando clasificadores bayesianos ingenuos (figura 23) empeora el desempeño con la experiencia y muestra alta varianza en ambos conjuntos de entrenamiento y test. Esto indica que los diferentes clasificadores dan una diferentes

persepectivas al problema resultando en la mejora de la métrica f2 al mismo tiempo que reduce la varianza en los resultados.

7. Resultados

7.1. Estabilidad de los Modelos

Finalmente se prueba evaluando repetidamente los clasificadores y ensambles sin variar las particiones del conjunto de datos para probar la estabilidad de los modelos finales, ya que este es un gran factor a la hora de elegir el mejor modelo posible para la naturaleza de los datos. Los resultados pueden ser observados en los diagramas de cajas en siguientes figuras:

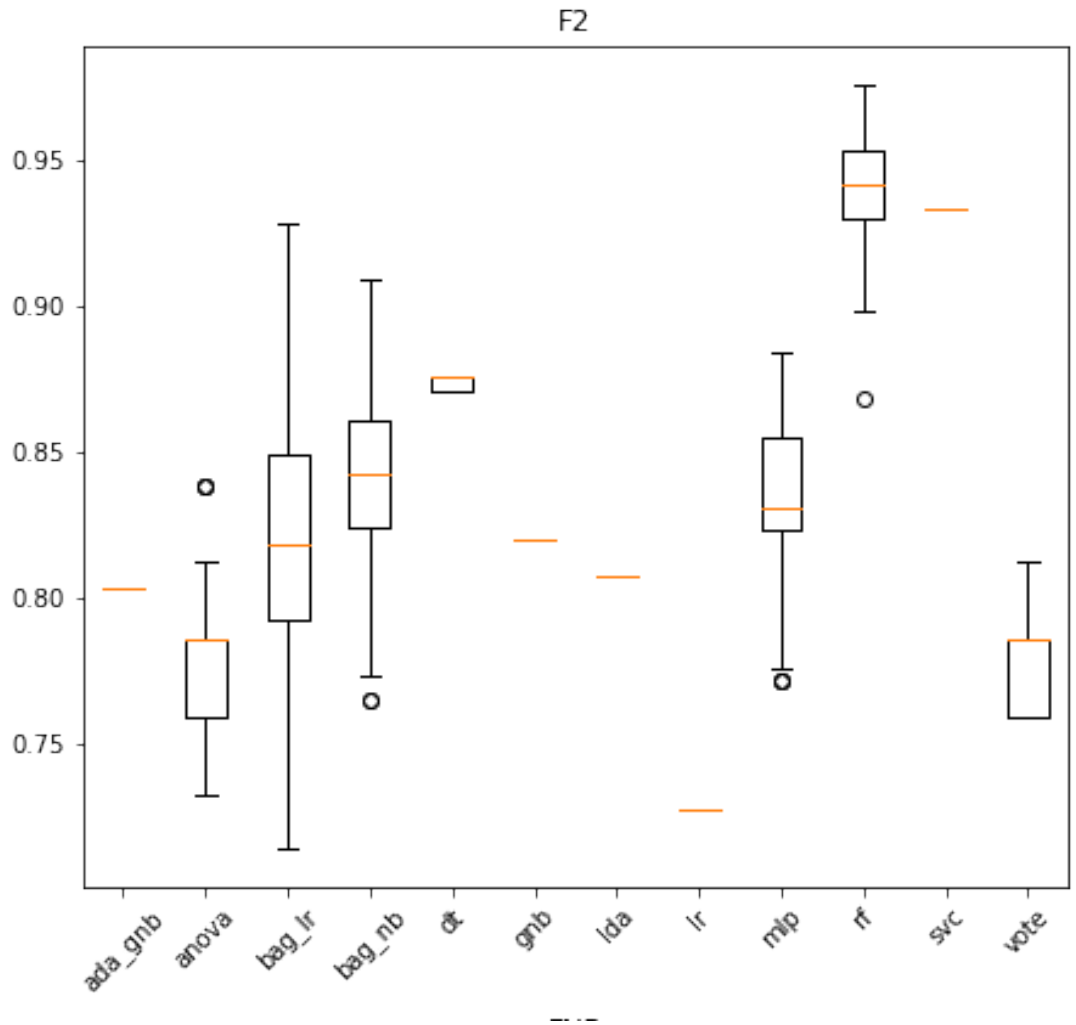

Figura 27: Comparación de estabilidad de los mejores modelos con la métrica f2

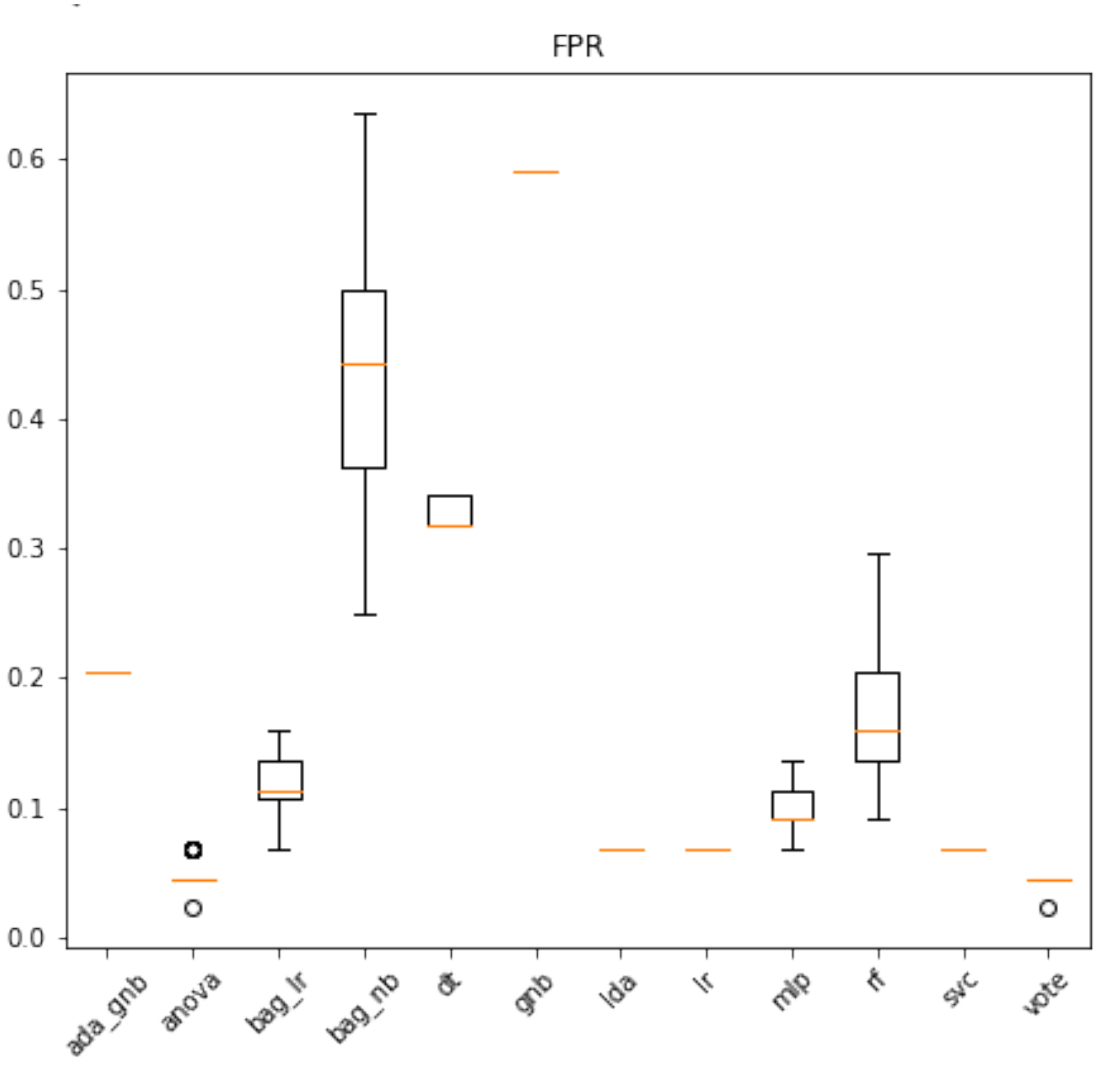

Figura 28: Resultados de tasa de falsos positivos.

Los métodos de perceptrón multicapa, bagging de regresión logística, y clasificador bayes ingenuo presentan menos estabilidad que los métodos de árboles de decisión, ensamble de máquinas de vectores de soporte y el mejor método de regresión logística. El ensamble de máquinas de soporte provee entonces una solución estable, con el mayor porcentaje de métrica f2 y garantiza reducir los falsos positivos a menos del 10 %.

7.2. Integrabilidad de la solución propuesta

La integración de las herramientas desarrolladas para la recolección de datos, procesamiento de gradientes y filtrado por medio del mejor modelo (ensamble de Máquinas de soporte vectorial) en un aeropuerto es fácilmente implementable en un entorno de estaciones GBAS. El presente proyecto fue financiado por PildoLabs para ser

* Master Thesis
** Facultad De Ciencias Básicas. Escuela de Física.
Maestría en Matemática Aplicada. Director: Raul Ramos Pollán.

integrado en el servicio de The CATalizer y actualmente se encuentra analizando datos de la región española y africana. Este servicio se hace en tiempo casi real pues el algoritmo tiene baja demanda computacional. Actualmente se encuentra entrenado y no se ha diseñado el flujo de trabajo para aumentar el conjunto de datos de entrada, pero en un futuro es recomendable entrenamiento continuo y con ello una estructura de manejo de datos a gran escala.

* Master Thesis
** Facultad De Ciencias Básicas. Escuela de Física.
Maestría en Matemática Aplicada. Director: Raul Ramos Pollán.

8. Conclusiones

- Se desarrolló un software para la adquisición de datos de receptores GNSS y su procesamiento para el cálculo de gradientes ionosféricos en GBAS, adicionalmente un módulo para la consolidación de estos datos en un conjunto de entrada para su análisis por medio de métodos de aprendizaje supervisado.
- Se obtuvo un modelo de machine learning para filtrar efectivamente falsas alarmas de los resultados de la metodología de cálculo de gradientes ionosféricos actualmente usada en aeropuertos.
- El algoritmo planteado es una solución para aumentar la disponibilidad de los sistemas GNSS en la industria de la aviación civil, es de bajo costo de instalación comparado con la adquisición de receptores de mayor presición o el diseño de un modelo del comportamiento de la ionósfera, además se toman en cuenta otros factores omitidos en los algoritmos en el estado del arte, los datos utilizados para el estudio son de diferentes regiones y diferentes épocas lo cual disminuye el sesgo del modelo actual.
- El tiempo de cómputo para probar el ensamble de máquinas de vectores de soporte en nuevos datos es proporcional al número de datos de entrada. La solución propuesta tiene la ventaja se aplicará en tiempo casi real, además se puede replicar y adaptar para aplicar a otras regiones.

* Master Thesis
** Facultad De Ciencias Básicas. Escuela de Física.
Maestría en Matemática Aplicada. Director: Raul Ramos Pollán.

Referencia Bibliográfica

[1] Akhoondzadeh, M. (2013). *Genetic algorithm for TEC seismo-ionospheric anomalies detection around the time of the Solomon (Mw=8.0) earthquake of 06 February 2013*. Advances in Space Research, 52(4), 581-59.

[2] Barrile, V., Cacciola, M., Morabito, F. C., y Versaci, M. (2006). *TEC Measurements Through GPS and Artificial Intelligence*. Journal of Electromagnetic Waves and Applications, 20(9), 1211-1220. https://doi.org/10.1163/156939306777442962

[3] Circiu, M.S et al.(2014) *Evaluation of Dual Frequency GBAS Performance Using Flight Data*. Proceedings of the 2014 International Technical Meeting of The Institute of Navigation: 645-656.

[4] Di Giovanni, G., y Radicella, S. M. (1990). *An analytical model of the electron density profile in the ionosphere*. Advances in Space Research, 10(11), 27-30. https://doi.org/10.1016/0273-1177(90)90301-F

[5] Encuentro ICAO ISTF/5. *GBAS Brazilian Ionospheric Threat Model. Project New Verification Methodology of Ionospheric Gradients Observed in the Brazilian Region. [fecha de consulta: 08 Noviembre del 2016]. Recuperado a partir de ¡*`http://www.icao.int/RO_APAC/Meetings/2015`*¿*

[6] Jung, S., Lee, J. (2012). *Long-term ionospheric anomaly monitoring for ground based augmentation systems.* Radio Science, 47(4). https://doi.org/10.1029/2012RS005016

[7] Klobuchar, J. A. (1987). *Ionospheric Time-Delay Algorithm for Single-Frequency GPS Users.* IEEE Transactions on Aerospace and Electronic Systems, AES-23(3), 325-331. https://doi.org/10.1109/TAES.1987.310829

[8] Lee, J., Jung, S., Bang, E., Pullen, S., Enge, P. (2010). *Long Term Monitoring of Ionospheric Anomalies to Support the Local Area Augmentation System* , Proceedings of the 23rd International Technical Meeting of The Satellite Division of the Institute of Navigation, Portland, OR, Septiembre del 2010, pp. 2651-2660.

[9] Lee, J., Yoon, M., Pullen, et al,. (2015)(pp. 1500-1506)*Preliminary Results from Ionospheric Threat Model Development to Support GBAS Operations in the Brazilian Region.* Presentado en 28th International Technical Meeting of The Satellite Division of the Institute of Navigation , Tampa, Florida

[10] Ma, G., y Maruyama, T. (2003). *Derivation of TEC and estimation of instrumental biases from GEONET in Japan.* Annales Geophysicae, 21(10), 2083-2093. https://doi.org/10.5194/angeo-21-2083-2003

[11] Ma X. F., Maruyama T., Ma G., y Takeda T. (2005). *Three-dimensional ionospheric tomography using observation data of GPS ground receivers and ionosonde by neural network*. Journal of Geophysical Research: Space Physics, 110(A5). https://doi.org/10.1029/2004JA010797

[12] Ordóñez C., Rodríguez-Pérez J., Moreira J., Matías J. M., y Sanz-Ablanedo E. (2011). *Machine Learning Techniques Applied to the Assessment of GPS Accuracy under the Forest Canopy.* Journal of Surveying Engineering, 137(4), 140-149. https://doi.org/10.1061/(ASCE)SU.1943-5428.0000049

[13] Sam, P., Young, P., y Enge, P. (2009).Impact and mitigation of ionospheric anomalies on ground-based augmentation of GNSS. *Impact and mitigation of ionospheric anomalies on ground-based augmentation of GNSS.* Radio Science, 44(1). https://doi.org/10.1029/2008RS004084

[14] Sánchez-Naranjo, S., Rincón, W., Ramos-Pollán, R., González, F. A., y Soley, S. (2017). A comprehensive assessment of ionospheric gradients observed in Ecuador during 2013 and 2014 for ground based augmentation systems. Advances in Space Research, 59(8), 1992-2006. https://doi.org/10.1016/j.asr.2017.02.001

[15] Sanz, J., Zornoza, J., & Hernández-Pajares, M. (2013). *GNSS DATA PROCESSING. Volume I: Fundamentals and Algorithms (Contactivity bv, Leiden, the Netherlands, Vol. 1). ESA Communications.*

[16] SINTEF. (2013). *Big Data, for better or worse: 90 % of world's data generated over last two years.* ScienceDaily. Recuperado en Abril 16 del 2018 a partir de www.sciencedaily.com/releases/2013/05/130522085217.htm

[17] Wang, C., Rosen, I. G., Tsurutani, B. T., Verkhoglyadova, O. P., Meng, X., y Mannucci, A. J. (2016). *Statistical characterization of ionosphere anomalies and their relationship to space weather events.* Journal of Space Weather and Space Climate, 6, A5. https://doi.org/10.1051/swsc/2015046

[18] Winston, P. (2010) *16. Learning: Support Vector Machines.* Recuperado a partir de https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-034-artificial-intelligence-fall-2010/.