



TESIS

MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DEL CONOCIMIENTO

**Pronóstico estadístico de precipitación mensual para la región de
Gran Chaco, Argentina.**

Juan Ignacio Mazza

Directora: Dra. Marcela Hebe González

Codirector: Dr. Julio Cesar Rodríguez Martino

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Agradecimientos

A mi mamá, a mi papá y a mi hermano, que me brindaron todo cuanto estuvo a su alcance para que pueda desarrollarme académica y profesionalmente.

A la Universidad de Buenos Aires y a la Facultad de Ciencias Exactas y Naturales por haberme brindado educación de excelencia, pública y gratuita.

A mis compañeros, que hicieron que los dos años de cursada de la maestría fueran una experiencia amena, compartieron conmigo su conocimiento, y me aconsejaron siempre que lo necesité

A Marcela González, por la generosidad de aceptarme como tesista, y por la paciencia con la que me explico y me guio a lo largo de este trabajo.

A Julio Rodríguez, no solo por aceptar dirigir y guiarme en esta tesis, sino por haberme brindado un lugar en un grupo de trabajo para que desarrolle mi carrera de científico de datos.

A Alfredo Rolla, por presentarme a Marcela y por sus aportes y sus consejos no solo para este trabajo sino durante los dos años de maestría.

A mis amigos y a mi pareja, que siempre me animaron a seguir y nunca me dejaron bajar los brazos.

Finalmente, al SMN, al INTA, y a la NOAA por la provisión de los datos utilizados para este trabajo

Tabla de contenidos

Agradecimientos.....	2
Índice de figuras.....	5
Índice de formulas	7
Índice de Tablas	8
Resumen.....	10
Abstract	10
CAPÍTULO I - INTRODUCCIÓN	
1.1. La precipitación en Argentina	11
1.2. Forzantes climáticos de la precipitación.....	13
1.3. Los modelos estadísticos	18
1.4. Aporte esperado al finalizar el proyecto	19
1.5. Transferencia de los resultados	20
1.6. Objetivos de la investigación	20
1.6.1. Objetivo general.....	20
1.6.2. Objetivos específicos.....	21
1.7. Hipótesis	21
CAPÍTULO II - DATOS	
2.1. Datos de variables meteorológicas y oceánicas a nivel global	23
2.2. Datos de precipitación del Gran Chaco Argentino.....	25
CAPÍTULO III - MÉTODOS Y RESULTADOS	
3.1. Regionalización del área de estudio.....	28
Grupo 0.....	31
Grupo 1.....	32
Grupo 2.....	33
Grupo 3.....	33
3.2. Determinación de Predictores	34
3.2.1. Análisis de Correlación	34
3.2.2. Definición de predictores	36
3.3. Diseño de modelos estadísticos.....	44
3.3.1. Redes neuronales artificiales.....	44
3.3.2. Regresión de Soporte Vectorial	46
3.3.3. XGBoost	49

Tesis - Juan Ignacio Mazza	
3.4. Ensamble de modelos.....	50
3.5. La eficiencia de los modelos de pronóstico.	55
3.6. Resultados obtenidos	58
3.6.1. El error de los modelos.....	58
3.6.2. La varianza de la precipitación explicada por los métodos	62
Grupo 0.....	62
Grupo 1.....	63
Grupo 2.....	64
Grupo 3.....	65
3.6.3. Los pronósticos categóricos.....	67
CAPÍTULO IV - CONCLUSIONES	
4.1. Conclusiones	74
4.2. Ideas para futuros desarrollos	76
4.3. Repositorio de archivos	76
4.4. Referencias	76
ANEXO 1 - Listado de estaciones y grupos	82
ANEXO 2 – Series temporales de precipitación mensual acumulada para cada estación (en mm) .	83
ANEXO 3 – Series medianas de precipitación acumulada (en mm) para cada grupo de estaciones	85
Grupo 0.....	85
Grupo 1.....	85
Grupo 2.....	87
Grupo 3.....	88
ANEXO 4 – Máscaras de filtrado.....	89
ANEXO 5 – Resultados de los modelos	92

Índice de figuras

Figura 1. Precipitación media estacional en Argentina (1981-2010).....	12
Figura 2. Onda anual de precipitación en Posadas y La Quiaca	12
Figura 3. Ciclo anual de precipitación en Angostura, Comahue y en Buenos Aires	13
Figura 4. Esquema representativo de una teleconexión.....	14
Figura 5. Fase cálida o El Niño y fase fría o La Niña sobre el océano Pacífico	15
Figura 6. Fase positiva del Dipolo del Indico.	16
Figura 7. Esquema de la fase positiva del SAM.	17
Figura 8. Esquema de la ubicación del anticiclón en el océano Atlántico sur	18
Figura 9. Estaciones meteorológicas de Gran Chaco, Argentina, y alrededores.	25
Figura 10. Neuronas de la red de Kohonen en la agrupación de estaciones meteorológicas	29
Figura 11. Estaciones meteorológicas agrupadas dibujadas sobre un mapa de la Argentina	30
Figura 12. Series temporales de precipitación mensual acum. para dos estaciones del grupo 0....	32
Figura 13. Series temporales de precipitación mensual acum. para dos estaciones del grupo 1....	32
Figura 14. Series temporales de precipitación mensual acum. para dos estaciones del grupo 2....	33
Figura 15. Series temporales de precipitación mensual acum. para dos estaciones del grupo 3....	34
Figura 16. Mapa de correlación entre las series de enero de HGT500 y febrero del grupo 0	35
Figura 17. Ejemplo de áreas de correlación significativa entre la variable SST en octubre y la precipitación de noviembre en el grupo 1	37
Figura 18. Máscara de filtrado para la variable HGT200.....	38
Figura 19. Máscara de filtrado para la variable U850.....	39
Figura 20. Ej. de aplicación de una máscara de filtrado para la variable HGT200.....	40
Figura 21. Ej. de aplicación de una máscara de filtrado con recorte de área para la var. U850.	40
Figura 22. Ejemplo de predictores para la precipitación de enero del grupo 1	42
Figura 23. Correlograma entre las variables predictoras y la clase para enero del grupo 1 [G1]	43
Figura 24. Ejemplo de red neuronal.....	44
Figura 25. Proyección de los vectores de entrada a un espacio de mayor dimensionalidad.....	47
Figura 26. Influencia del parámetro 'C' en el hiperplano de separación de una SVM.....	48
Figura 27. Predicción individual del modelo 'Red neuronal' para la precip. de enero del G1.	52
Figura 28. Predicción individual del modelo SVR para la precip. de enero del G1.	52
Figura 29. Predicción individual del modelo 'XGBoost' para la precipitación de enero del G1.	53
Figura 30. Predicción generada por el ensamble de los modelos para la precip. de enero del G1..	53
Figura 31. Ejemplo de curvas ROC y sus respectivos AUC.....	56

Figura 32. MAE y varianza explicadas del ensamble (M01 - G1)	57
Figura 33. Tabla de precisión, recall y F1-score para el ensamble (M01 - G1).....	57
Figura 34. AUC para el modelo de ensamble (M01 - G1).....	58
Figura 35. Errores absolutos medios (MAE) para cada grupo-mes	59
Figura 36. MAEs relativizados por la precipitación mediana de cada Grupo-Mes	61

Índice de formulas

Fórmula 1. Int. de conf usado en el cálculo de una correlación significativa con 95% de conf.	36
Fórmula 2. Error absoluto medio (MAE por sus siglas en ingles)	50
Fórmula 3. Varianza explicada.....	50
Fórmula 4. Formula de ponderación de las predicciones.	51
Fórmula 5. Precision	55
Fórmula 6. Recall	56
Fórmula 7. F1-Score	56

Índice de Tablas

Tabla 1. Errores absolutos medios para cada par Grupo-Mes	60
Tabla 2. Medianas de precipitación para cada par Grupo-Mes	60
Tabla 3. MAEs relativizados por la precipitación mediana de cada Grupo-Mes	60
Tabla 4. Varianzas explicadas de los modelos individuales del grupo 0	63
Tabla 5. Varianzas explicadas de los modelos individuales del grupo 1	64
Tabla 6. Varianzas explicadas de los modelos individuales del grupo 2	65
Tabla 7. Varianzas explicadas de los modelos individuales del grupo 3	66
Tabla 8. Curvas AUC de los ensambles para todas las categorías de todos los grupos	68
Tabla 9. Estadísticos Precision, Recall y F1-Score para el modelo ensamble del grupo 0	69
Tabla 10. Estadísticos Precision, Recall y F1-Score para el modelo ensamble del grupo 1	70
Tabla 11. Estadísticos Precision, Recall y F1-Score para el modelo ensamble del grupo 2	71
Tabla 12. Estadísticos Precision, Recall y F1-Score para el modelo ensamble del grupo 3	72

CAPÍTULO I

INTRODUCCIÓN

Resumen

Disponer con antelación del pronóstico de lluvias para una región resulta importante a la hora de planificar y ejecutar actividades que podrían verse afectadas por estas, con el fin de minimizar o mitigar los riesgos asociados a posibles temporadas de sequía o inundación.

En la actualidad el pronóstico de lluvias para la región de Gran Chaco, Argentina, es muy eficiente a corto plazo (hasta 10 días), pero su performance disminuye para escalas superiores al mes. En regiones localizadas, una forma de abordar el pronóstico en estas escalas más grandes es la utilización de métodos estadísticos que logran aprender del pasado y generan modelos de predicción a futuro. Para esto, han sido utilizados modelos de regresión lineal múltiple que combinan varias variables atmosféricas y oceánicas. Sin embargo, dada la naturaleza no lineal y caótica de la lluvia otras metodologías podrían mejorar dichos pronósticos.

El presente trabajo propone explorar diversos modelos no lineales de **machine learning** para pronosticar las lluvias mensuales en la región de Gran Chaco, generando una solución automatizada aplicable para cualquier región de la Argentina adaptando el conjunto de predictores.

Abstract

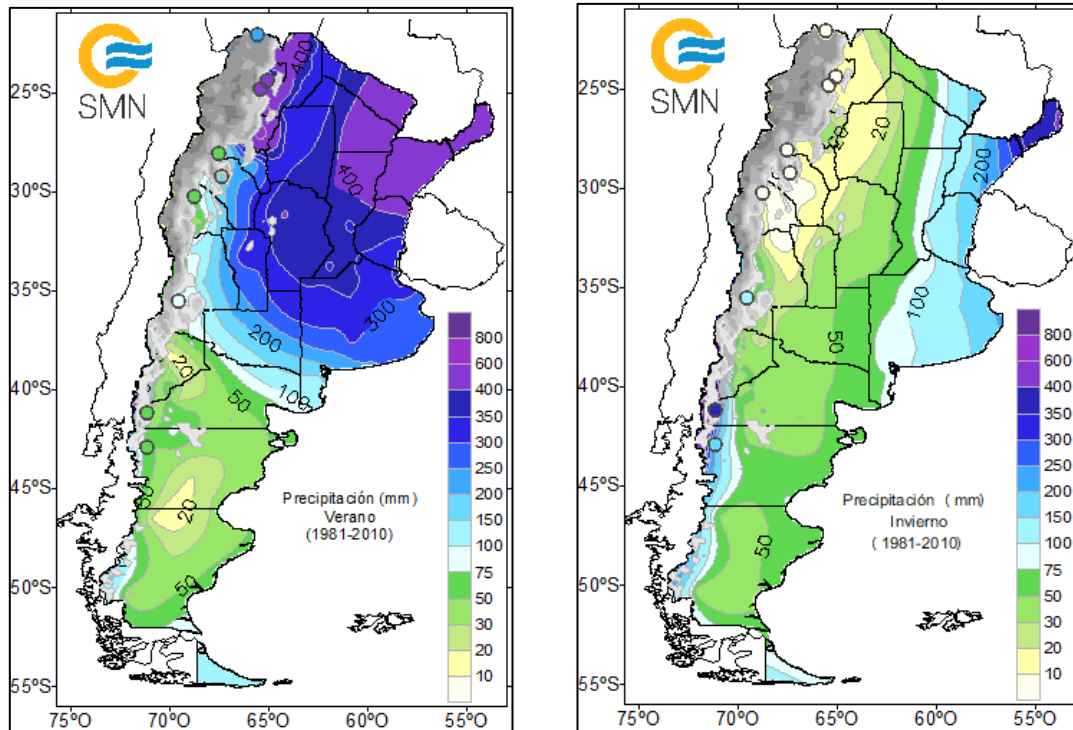
Having the rain forecast for a region in advance is important when planning and executing activities that could be affected by them, in order to minimize or mitigate the risks associated with possible drought or flood seasons.

Currently, the rain forecast for the Gran Chaco region, Argentina, is very efficient in the short term (up to 10 days), but its performance decreases for scales greater than one month. In localized regions, one way to approach forecasting at these larger scales is to use statistical methods that learn from the past and generate future prediction models. For this, multiple linear regression models have been used that combine various atmospheric and oceanic variables. However, given the non-linear and chaotic nature of rainfall [1], other methodologies could improve such forecasts.

The present work proposes to explore various non-linear models of machine learning to forecast monthly rains in the Gran Chaco region, generating an automated solution applicable to any region of Argentina by adapting the set of predictors.

1.1. La precipitación en Argentina

Argentina posee diferentes regímenes de precipitación [2], [3]. En la figura 1 se muestran las isoyetas (líneas que unen puntos de igual precipitación) de precipitación acumulada en las diferentes estaciones del año, calculadas en base al período 1981-2010 (Fuente: Servicio Meteorológico Nacional). Estos valores son promedios en el período para la precipitación acumulada durante toda la estación del año.



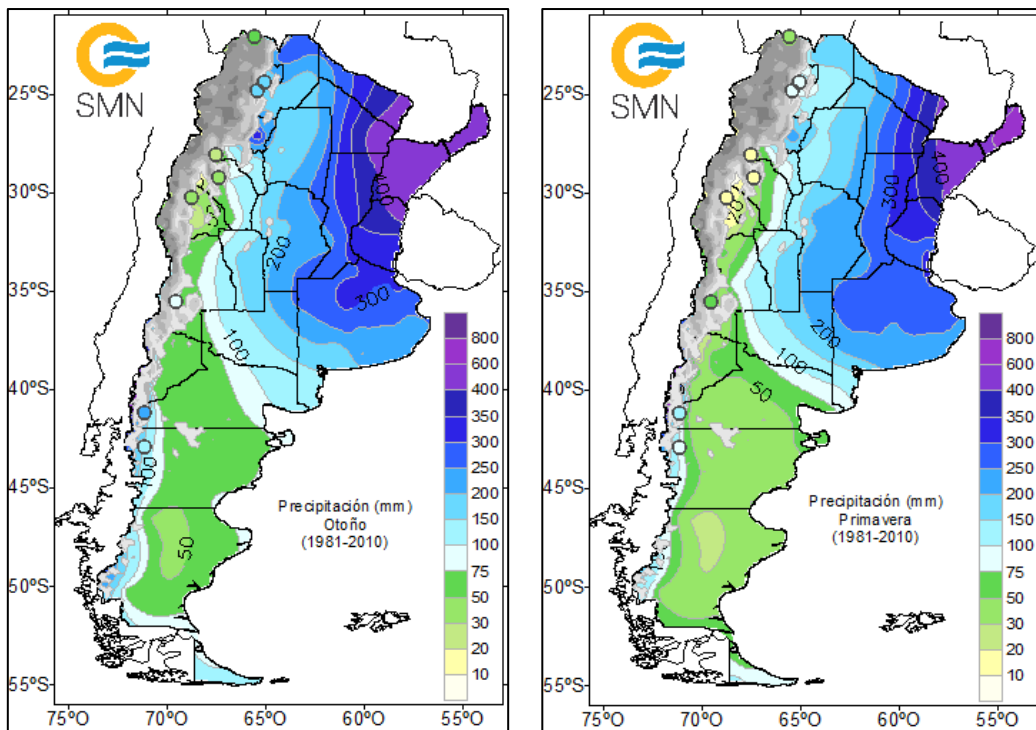


Figura 1. Precipitación media estacional en Argentina (1981-2010). Fuente: Servicio Meteorológico Nacional.

La región subtropical, en el noreste de Argentina tiene precipitación alta durante todo el año, en verano los acumulados superan los 400mm y en invierno están entre 200 y 300mm. La figura 2 muestra el ejemplo de Posadas. Hacia el oeste, se empieza a producir un ciclo anual marcado con mínimos en invierno, de modo tal que el noroeste argentino existe una gran diferencia entre la alta precipitación acumulada estival (entre 200 y 400mm) y el invierno que prácticamente no tiene precipitación (menos de 20mm). La figura muestra el ciclo anual de precipitación en La Quiaca. El Gran Chaco Argentino se ubica en esta región de Argentina subtropical con grandes diferencias en el régimen de precipitación entre los sectores este y oeste.

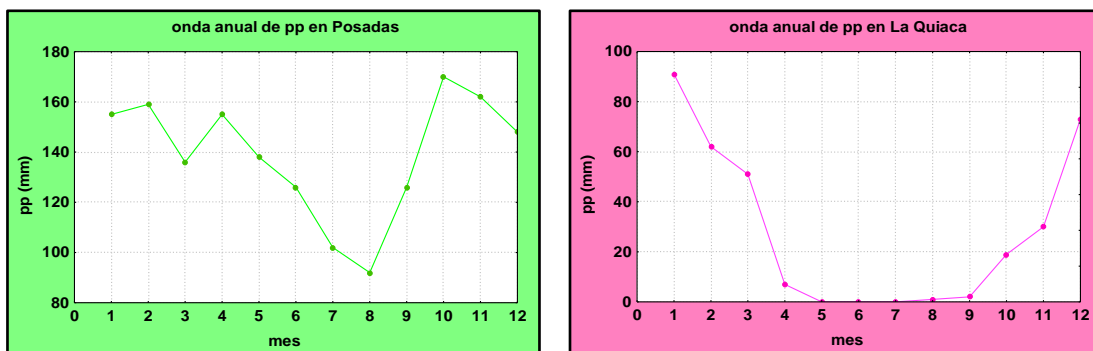


Figura 2. Onda anual de precipitación en Posadas (panel izquierdo) y La Quiaca (panel derecho)

La zona de la cordillera de Los Andes tiene comportamientos diversos. Al norte de 38°S es alta y maciza con valles en dirección norte-sur, lo que impide el ingreso de aire húmedo desde el océano Pacífico. Esto genera áreas muy secas como puede observarse en el caso de Cuyo. Al sur de 38°S, la cordillera es más baja, menos maciza y los valles presentan dirección este-oeste. El ingreso de aire desde el Pacífico es posible, lo que produce que la zona del noroeste patagónico, conocida con Comahue, presente un ciclo anual de lluvias con máximos en invierno (superando los 200mm). A modo de ejemplo la figura 3 muestra la precipitación media mensual en Angostura, una estación de alta montaña en la zona de Comahue. El sur de la cordillera de Los Andes, en la zona sur de Santa Cruz y Tierra del Fuego tienen un régimen de precipitación con lluvias durante todo el año (alrededor de 150mm), asociadas a los sistemas frontales de las bajas presiones ubicadas en esas latitudes.

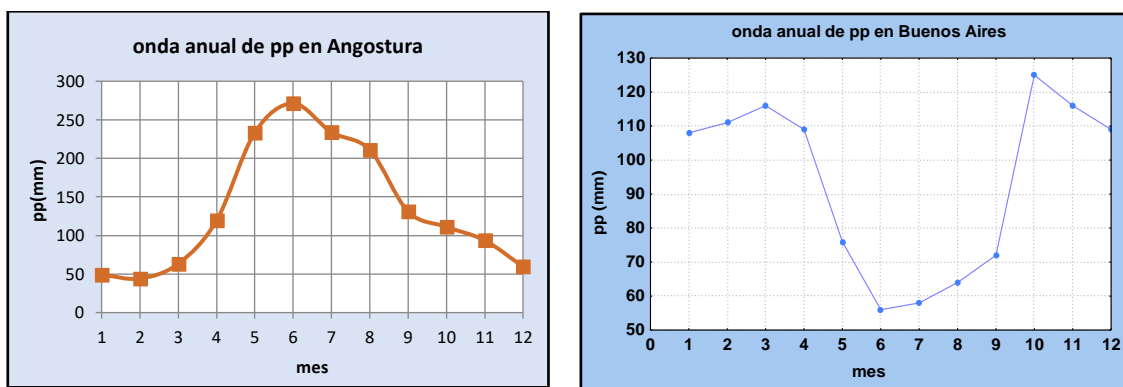


Figura 3. Ciclo anual de precipitación en Angostura, Comahue (panel izquierdo) y en Buenos Aires (panel derecho)

Por otro lado, la Patagonia este es una meseta con baja precipitación durante todo el año (debajo de los 75mm). En el centro y este de Argentina el régimen de lluvias es tal que la precipitación estival supera a la invernal, pero se producen dos máximos relativos en las estaciones de transición (otoño y primavera). En la figura 3 se muestra el ejemplo de la precipitación en Buenos Aires (panel derecho) con máximos en marzo y octubre.

1.2. Forzantes climáticos de la precipitación

La importancia de tener un pronóstico extendido de precipitaciones en escalas mensuales radica en que pueda ser utilizado como una herramienta para los tomadores de decisión en diferentes actividades, como la programación anticipada del manejo de presas para la generación de energía hidroeléctrica, el mejor manejo de la agricultura, la ganadería y la prevención de desastres como incendios por sequías o grandes inundaciones. Si bien los modelos meteorológicos han logrado una gran eficacia en pronósticos del tiempo hasta los 10 días, la predicción en escalas mensuales y

estacionales aún tiene muchas deficiencias y es un área actualmente en constante desarrollo [4]–[7]. En particular, en Sudamérica algunos autores han evaluado la eficiencia del pronóstico estacional y coinciden en que aún hace falta avanzar en la investigación para conseguir mejores resultados [8]–[10].

Para realizar predicciones en escalas medianas se pueden utilizar modelos determinísticos o estadísticos. Los primeros consisten en la resolución de un sistema de ecuaciones que simulan la física atmosférica en forma aproximada. Los modelos estadísticos se basan en el aprendizaje de lo ocurrido en el pasado. El principio fundamental se basa en que las variaciones lentas de las condiciones de los océanos (temperatura de la superficie marítima) y/o de la atmósfera (vientos, presión atmosférica medida a diferentes alturas, cantidad de agua precipitable en el aire, entre otras) pueden inducir cambios en la circulación de la atmósfera y por lo tanto en la precipitación. Dentro de dichos fenómenos de lenta variación se deben destacar las llamadas “teleconexiones” (conexión a la distancia, Figura 4) [11]–[16]. Su definición se basa en que una anomalía que se produce en una variable meteorológica en una región del planeta, activa trenes de onda que producen anomalías de otras variables en una región lejana.



Figura 4. Esquema representativo de una teleconexión

Dentro de las teleconexiones más estudiadas se encuentra el fenómeno El Niño-Oscilación Sur (ENOS) [17][18][19][20] que presenta una respuesta muy importante en la precipitación de Argentina, sobre todo en la zona subtropical. En particular la fase cálida del ENOS, conocida como “El Niño” y que corresponde a un calentamiento anómalo de océano Pacífico central ecuatorial, se relaciona con anomalías positivas de precipitación en el noreste de Argentina y en el Comahue durante la primavera [21], [22]. La fase fría del ENOS (“La Niña”) en general se asocia a situaciones más secas. La figura 5 muestra un esquema de ambas fases del ENOS. Es un fenómeno que presenta una gran variabilidad, que no se produce regularmente en el tiempo y las consecuencias son muy dependientes del área y de la época del año. La definición de una fase cálida (fría) se produce cuando las anomalías de la temperatura de la superficie del mar el Pacífico ecuatorial superan los 0,5°C

(están debajo de los $-0,5^{\circ}\text{C}$) y se mantienen al menos 6 meses en esos valores. Por lo tanto, resultan condiciones bastante excepcionales y en la mayoría de los casos no se cumplen, lo que define una situación de ENOS neutral.

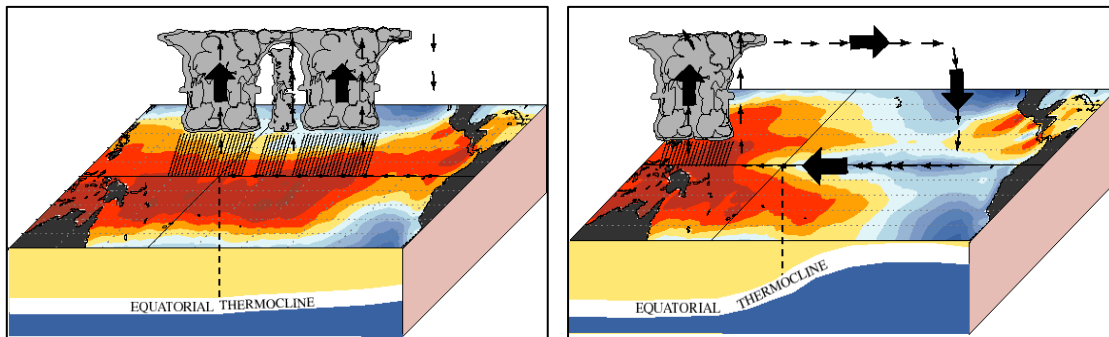


Figura 5. Fase cálida o El Niño (panel izquierdo) y fase fría o La Niña (panel derecho) sobre el océano Pacífico.

Las zonas naranjas de la figura 5 se corresponden con anomalías positivas y las zonas azules con anomalías negativas de la temperatura de la superficie del mar. Fuente: [23]

Existen otros forzantes de teleconexión que actúan paralelamente al ENOS y que adquieren relevancia sobre todo en situaciones de ENOS neutrales. Existe una clara influencia de las anomalías de temperatura de la superficie del mar en el océano Índico sobre la precipitación, ya sea a través del patrón del dipolo del Índico [24][25] (figura 6) como de su calentamiento o enfriamiento generalizado [15], [26], [27]. La fase positiva del dipolo del Índico se define como un enfriamiento del noreste del Índico y un calentamiento del sudoeste, la fase negativa se define con el comportamiento opuesto. Por otro lado, un calentamiento o enfriamiento generalizado de toda la cuenca del Índico, también se relaciona con cambios en la precipitación de la región de Sudamérica. Varios autores han abordado este tema, por ejemplo, Zheng y Frederiksen, han logrado demostrar que existe una fuerte correlación entre la temperatura de la superficie del océano Índico y las lluvias de invierno y verano en Nueva Zelanda. [11] C. J. C. Reason, demostró como las lluvias de Sudafrica se relacionan con cambios de temperatura en la superficie del océano Índico. [28] Gissila T. et al. demostraron la relación que existe entre la temperatura de la superficie del océano Índico y las lluvias de verano en Etiopía. Asimismo, [25] Chan et al, comprobaron que existe una relación directa entre el dipolo que se genera en el océano Índico y las precipitaciones en la Cuenca del Plata y el sur de Brasil.

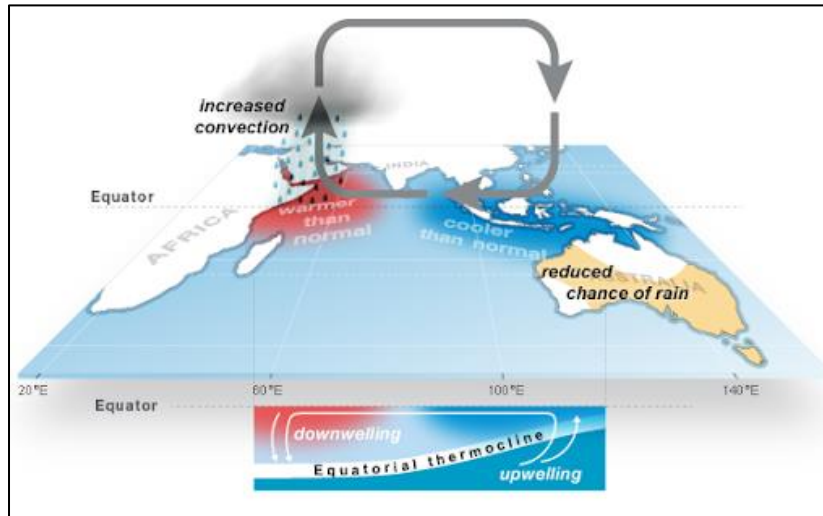


Figura 6. Fase positiva del Dipolo del Índico. Fuente: [29]

Existen otros forzantes más regionales que afectan a la precipitación. Uno de ellos es el llamado “Monzón Sudamericano” [30][31][32]. Durante el verano austral la convergencia intertropical, que es una zona convectiva, con gran nubosidad y lluvia sobre el Ecuador, se desplaza hacia el sur. En las zonas continentales encuentra una fuente de energía dado que la tierra se calienta diferencialmente al océano y por lo tanto una lengua convectiva se desplaza sobre Sudamérica. Esto hace que en verano se produzcan fuertes precipitaciones en la selva brasilera. Por otro lado, la presencia del anticiclón semipermanente del océano Atlántico genera una circulación de aire que proviene de este y se introduce en Sudamérica, encuentra una cordillera de Los Andes que lo hace desviar la circulación y la canaliza hacia el sur. En verano, este proceso se produce sobre la selva brasilera lluviosa y por lo tanto el norte y centro de Argentina reciben un flujo de aire muy húmedo desde el norte que genera condiciones propicias para la precipitación [33][34]. Por lo que el noroeste y centro de Argentina tienen un régimen de precipitación con lluvias predominantemente estivales.

Otra característica que se relaciona muy fuertemente con la precipitación en Argentina son los sistemas precipitantes de tipo frontal que se desplazan desde el sur hacia el noreste. Los trenes de ondas de Rossby [21] [22] se desplazan sobre el océano Pacífico y acceden al territorio argentino al sur de 38°S, donde la cordillera de los Andes es más baja y se desplazan hacia el noreste trayendo aire frío polar hacia las bajas latitudes. El aire frío desplaza al aire subtropical generando lo que se conoce como “frente frío” que tiene asociado inestabilidad y precipitación. Asociado a estos sistemas se define el llamado: “modo anular del Sur” (SAM, por sus siglas en inglés) [35] (Figura 7), su fase positiva se produce cuando la presión en la zona polar y en el cinturón de bajas presiones subpolares (alrededor de los 50-60°S) y el de altas presiones subtropicales (alrededor de 30°S) están

intensificados. Esto genera fuertes vientos del oeste en latitudes medias que retardan el pasaje de los frentes hacia el norte y por lo tanto se relacionan con menor precipitación en Argentina [36][37][38]. La fase negativa se define en forma opuesta, favorece el pasaje de los frentes y por lo tanto la precipitación en Argentina.

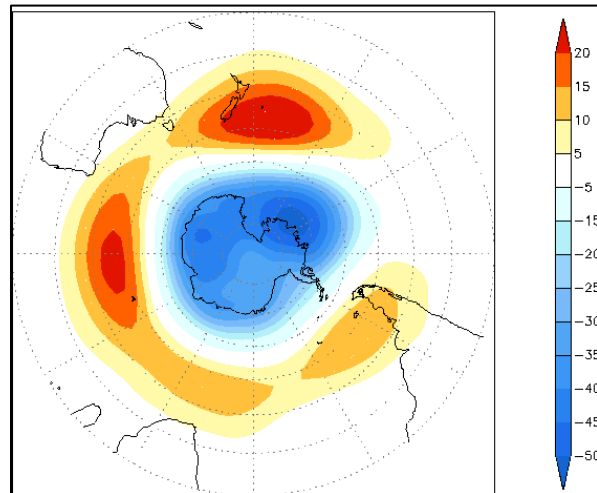


Figura 7. Esquema de la fase positiva del SAM. La zona celeste implica bajas presiones y las rojas altas presiones, ambas intensificadas. Fuente: Thompson y Wallace (2000)

Por otro lado, la posición e intensidad del anticiclón semi-permanente del Atlántico también regula los procesos advectivos de humedad hacia el continente proveniente de dicho océano, que influyen la precipitación en el norte argentino [39][40]. El anticiclón del Atlántico sur se ubica alrededor de los 30°S sobre el océano, gira en forma antihoraria y siempre se desplaza hacia el hemisferio donde es verano. La figura 8 muestra un esquema de la posición del anticiclón en el océano Atlántico. El aire cálido y húmedo ingresa al continente sudamericano a través de dicho anticiclón, con intensidad y posición que depende de muchos factores y por lo tanto afecta la humedad que recibe el continente y a la posibilidad de generar precipitación. En Argentina, muchos autores han abordado esta problemática. Por ejemplo, [41] Oliveri P., detectó en su tesis de licenciatura que la advección de calor y humedad provenientes de los anticiclones semi-permanentes del Atlántico Sur y del Pacífico Sur, influyen de forma relevante sobre la temperatura y la precipitación en escalas estacionales en Argentina con una mayor señal para la temperatura en invierno. Una detallada revisión sobre la influencia del anticiclón del Atlántico sobre la precipitación en Argentina fue realizada por Garbarini et al. [40]

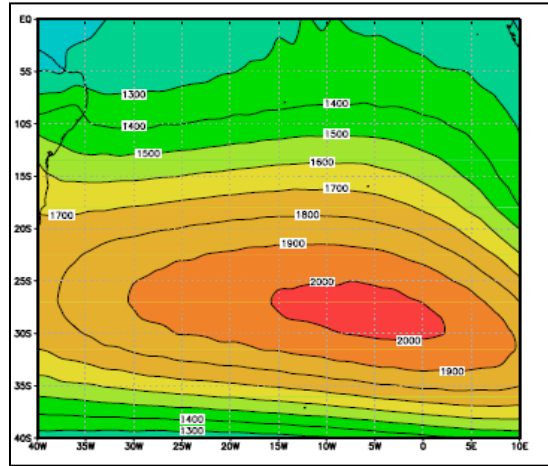


Figura 8. Esquema de la ubicación del anticiclón en el océano Atlántico sur (entre el ecuador y 40°S y entre 40°W y 10°E). La zona roja es la que presenta mayor presión atmosférica en la superficie.

Todos estos factores afectan la variabilidad interanual de la precipitación en Argentina, por lo cual pueden ser considerados forzantes climáticos de la misma y servirán para la para la definición de predictores de modelos estadísticos [8].

1.3. Los modelos estadísticos

Los modelos estadísticos presentan varias incertezas. Por un lado la atmósfera tiene una componente caótica que no es predecible [1], pero además existen otras fuentes de error: una relacionada con la definición de predictores correctos y los conjuntos de datos con los cuales se construyen y otra asociada a las metodologías estadísticas que se utilizan para generar los modelos [42].

Es aquí donde entra en juego el data mining, un area de estudio interdisciplinario que analiza grandes volúmenes de datos en busca de patrones relevantes, útiles y previamente desconocidos.

Estas técnicas computacionales aún no han recibido mucha atención en la comunidad meteorológica, en general, sus bondades no son del todo utilizadas y sin embargo, son realmente útiles para abordar diferentes tipos de pronósticos [43]. En general sólo se han utilizado redes neuronales sencillas y la mayoría del trabajo sobre interpretación de redes neuronales en meteorología se ha centrado hasta ahora en tareas de clasificación de imágenes satelitales. Uphoff y Hilburn (2020) detallan muchas estrategias y consideraciones prácticas para la utilización de técnicas de data mining en aplicaciones meteorológicas, incluido los métodos potenciales para la

Tesis - Juan Ignacio Mazza

interpretación con experimentos sintéticos y mapas de calor emergiendo como herramientas particularmente útiles. Los meteorólogos tienen un papel crucial que desempeñar planteando situaciones que puedan ser resueltas con estos métodos lo que contribuirá en gran medida a crear redes neuronales confiables para uso operativo en meteorología.

Utilizando datos de variables atmosféricas y oceánicas a nivel global, junto con datos de precipitación de las estaciones meteorológicas de la región de Gran Chaco se pretende aplicar diferentes técnicas de data mining para recolectar y procesar dichos datos a fines de generar un conjunto de datos final que pueda ser utilizado para entrenar diferentes técnicas como redes neuronales artificiales, regresión de soporte vectorial y XGBoost con el objetivo de predecir la precipitación de una región en un determinado mes, en otras palabras, aprender del pasado para predecir el futuro. Esta forma de predicción de la precipitación se diferencia del modelo determinístico tradicional donde dadas ciertas condiciones iniciales aplicadas a un sistema de ecuaciones que rigen la atmósfera, se realizan simulaciones para obtener un pronóstico. Tanto los modelos estadísticos como los determinísticos tienen una eficiencia limitada en escalas mensuales. Es por ello que las técnicas de data mining presentan una alternativa para explorar si dicha eficiencia puede ser mejorada.

Los modelos de predicción propuestos en este trabajo abordan esta problemática desde el punto de vista de la utilización de técnicas de data mining, procesando una gran cantidad de datos históricos y generando predicciones de precipitación mensual para la zona del Gran Chaco Argentino.

1.4. Aporte esperado al finalizar el proyecto

El aporte de esta tesis de maestría es la posibilidad de contar con modelos no lineales que puedan predecir la precipitación con un mes de antelación en diferentes regiones del Gran Chaco argentino. Actualmente el pronóstico en escalas mensuales está poco desarrollado. Los modelos dinámicos, o sea basados en ecuaciones para la atmósfera, tienen en general poca eficiencia para predecir la lluvia [44][45][46] y, por lo tanto, los métodos estadísticos complementan estos pronósticos modelando la precipitación a través del conocimiento de su comportamiento en el pasado. En este trabajo, se deriva un esquema mensual de pronóstico de precipitación para la región del Gran Chaco utilizando metodologías estadísticas no lineales.

1.5. Transferencia de los resultados

El grupo de trabajo que dirige la Dra. Marcela H. González (codirectora de esta tesis), está consolidado y trabaja conjuntamente con el Servicio Meteorológico Nacional (SMN) en la elaboración de los pronósticos trimestrales de precipitación y temperatura cada mes en reuniones de consenso realizadas para tal fin desde 2007 (actualmente se encuentra un convenio de colaboración en trámite). Además, ha firmado acuerdos específicos como, por ejemplo: en diciembre de 2015 por resolución N°3174 del CD de la FCEN UBA, un Convenio Marco de Cooperación Académica entre la Facultad de Ingeniería de la Universidad del Comahue y la FCEN UBA donde este grupo de trabajo se compromete a la cooperación mutua para el desarrollo de modelos estadísticos de pronóstico de precipitación y temperatura para la región del Comahue, con el objetivo de mejorar la operatividad de las presas hidroeléctricas o una carta de intención de cooperación mutua con la Comisión Regional del Río Bermejo (COREBE) firmada en setiembre de 2008 entre CIMA (CONICET-UBA) y COREBE. Una muestra de la labor del grupo dentro de las actividades del Departamento de Ciencias de la Atmósfera y los Océanos (FCEN-UBA) puede visualizarse en su página web: <http://perspectiva.at.fcen.uba.ar>

Esto prueba que todos los resultados de de esta tesis de maestría pueden ser transferidos y utilizados por instituciones nacionales y provinciales para mejorar los pronósticos regionales y orientar a las actividades que dependen altamente del clima.

1.6. Objetivos de la investigación

1.6.1. Objetivo general

Valiéndose de datos meteorológicos de dominio público (Ver sección 2.1), esta tesis de maestría plantea demostrar que es posible utilizar varias variables atmosféricas y oceánicas para predecir a mediano plazo (pronóstico mensual) la precipitación acumulada para la región de Gran Chaco en Argentina. El modelo utilizará los datos de las variables de reanálisis (archivos NC) de un mes y utilizarlos para predecir las lluvias de la región del mes siguiente.

1.6.2. Objetivos específicos

1. Regionalizar el área de estudio del Gran Chaco Argentino
2. Definir predictores para la precipitación en cada mes del año y en cada región
3. Elaborar modelos de predicción estadística para cada mes del año
4. Determinar la performance de dichos modelos

1.7. Hipótesis

Es posible desarrollar un modelo de machine learning que encuentre predictores entre las variables de superficie terrestre y atmósfera que puedan ser usadas para generar modelos no lineales para pronosticar a mediano plazo (un mes) las precipitaciones de la región de Gran Chaco

CAPÍTULO II

DATOS UTILIZADOS

Este capítulo de la tesis explica en detalle los datos utilizados: Los organismos públicos que los administran y disponibilizan, su descarga automática, la preparación y explotación de los mismos.

2.1. Datos de variables meteorológicas y oceánicas a nivel global

El proyecto NCEP/NCAR reanalysis [47] es un proyecto de Physical Science Laboratory de la National Oceanic and Atmospheric Administration (NOAA) y la Universidad de Colorado, con el apoyo del Departamento de Energía de Estados Unidos, para recolectar datos durante los siglos XX y XIX, combinarlos con registros de satélite y radar y así generar como producto final una base de datos en puntos de retículo con resolución de 2,5°. Actualmente cuentan con datos históricos desde el año 1948 hasta la actualidad. Sin embargo, los datos previos a 1979 no tienen incorporada la información proveniente de satélites por lo cual resultan menos confiables. Por esta razón en general, se utilizan los datos posteriores a 1979. Estos reanálisis asimilan solo las observaciones superficiales de la presión en el Sistema de Pronóstico Global de la NOAA y prescriben la temperatura de la superficie del mar y la distribución del hielo marino para estimar, por ejemplo, la temperatura, la presión, los vientos, la humedad, la radiación solar y las nubes, desde la superficie hasta la parte superior de la atmósfera. Utiliza un método de asimilación de datos de filtro de conjunto que estima directamente el estado más probable de la atmósfera global para cada período de tres horas y también estima la incertidumbre en ese análisis. El proyecto ha contado con la cooperación internacional en el marco de la iniciativa de Reconstrucciones de la Circulación Atmosférica sobre la Tierra (ACRE), que emprende y facilita la recuperación de observaciones meteorológicas globales terrestres y marinas instrumentales históricas para respaldar las reconstrucciones meteorológicas de los últimos 200 años. El Sistema Mundial de Observación del Clima y el Programa Mundial de Investigación del Clima de la Organización Meteorológica Mundial han brindado apoyo adicional.

En este trabajo se utilizaron las siguientes variables meteorológicas:

1. Altura geopotencial en diferentes niveles de presión
 - a. 200 hp o niveles altos (HGT200)
 - b. 500 hp o niveles medios (HGT500)
 - c. 1000 hp o niveles bajos (HGT1000)
2. Temperatura de la superficie del mar (SST)
3. Cantidad de agua precipitable en la columna atmosférica (TCW)
4. Componente zonal del viento (dirección Oeste-Este) en capas bajas (U850)
5. Componente meridional del viento (dirección Sur-Norte) (V850)

Cada uno de los siete archivos mencionados contiene los datos de la variable que representa en una matriz de tres dimensiones:

1. **Latitud:** Esta se define como la medida angular de la distancia entre un punto y el Ecuador, sea hacia el norte o hacia el sur. Esta dimensión de los datos de entrada cuenta con 73 valores de latitud
2. **Longitud:** Se define como la medida angular de la distancia entre un punto y el Meridiano de Greenwich, sea hacia el este o hacia el oeste. Esta dimensión de los datos de entrada está dividida en 144 valores de longitud.
3. **Tiempo:** Una medición diaria entre el 01-01-1948 y el 31-12-2019, los datos anteriores a 1979 no serán tenidos en cuenta, debido a que sólo los datos desde 1979 al presente son confiables gracias a la utilización de satélites espaciales.

La descarga de estos datos se realiza de manera automática mediante un script en Python que se conecta al servidor NCEP/NCAR reanalysis utilizando el protocolo FTP (URL: <ftp://ftp.cdc.noaa.gov/Datasets/ncep.reanalysis.derived/pressure>) y descarga los siguientes archivos:

- hgt.mon.mean.nc
- uwnd.mon.mean.nc
- vwnd.mon.mean.nc
- pr_wtr.mon.mean.nc
- sst.mnmean.nc

Cabe destacar que el archivo **hgt.mon.mean.nc** contiene los valores de presión atmosférica a diferentes altitudes, para esto, además de las dimensiones **latitud**, **longitud** y **tiempo**, utiliza una cuarta dimensión llamada **level**. Como el presente trabajo sólo utiliza los valores de presión atmosférica en tres altitudes específicas (200, 500 y 1000) esta cuarta dimensión es removida por el script de descarga generando en cambio un archivo específico para cada altitud: **hgt200.nc**, **hgt500.nc** y **hgt1000.nc**

2.2. Datos de precipitación del Gran Chaco Argentino

Se utilizaron registros de precipitación mensual provenientes de 34 estaciones meteorológicas correspondientes al área de Gran Chaco Argentino (área verde) y sus alrededores (área gris) (Figura 9) para el período 1979-2019. La fuente de datos es la red de medición del Servicio Meteorológico Nacional (<https://www.smn.gov.ar/>) y del Instituto Nacional de Tecnología Agropecuaria (<https://www.argentina.gov.ar/inta>). Las estaciones cercanas a la región de Gran Chaco se incluyen con el objeto de obtener mayor precisión en la zona de frontera de la región. La figura 1 muestra la localización de cada una de las estaciones utilizadas en el estudio que comprende su ubicación en latitud y longitud y la altura de la misma. La tabla en el ANEXO 1 muestra la longitud, latitud y elevación de las estaciones utilizadas.

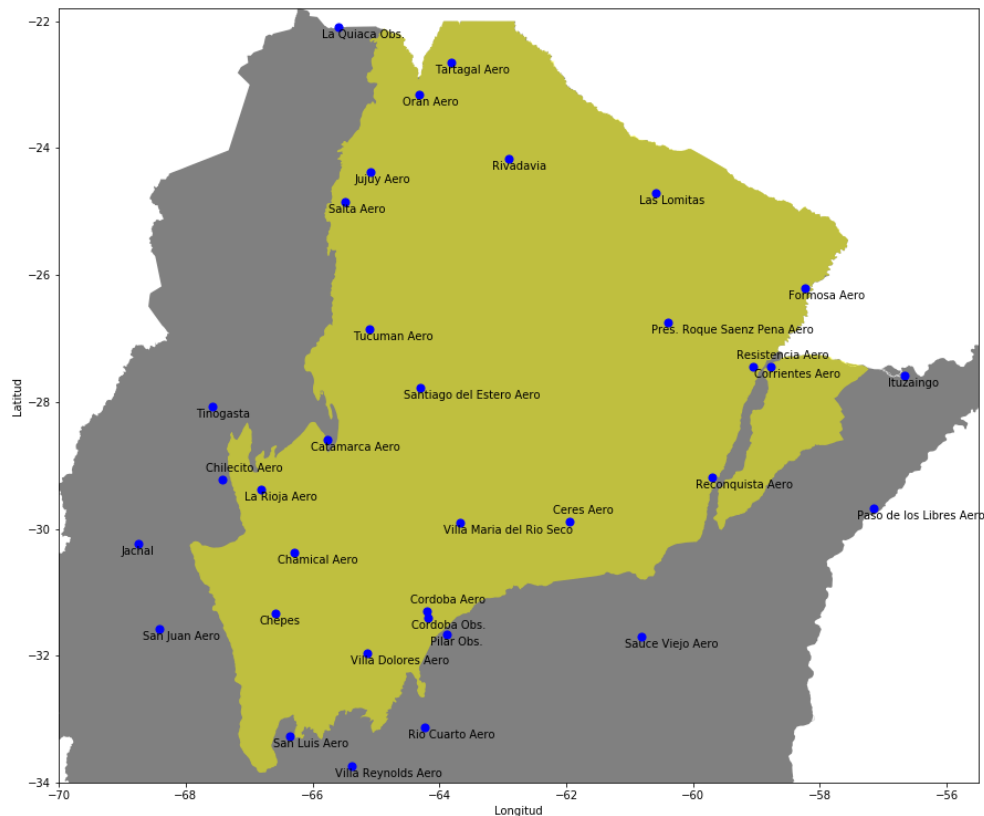


Figura 9. Estaciones meteorológicas de Gran Chaco, Argentina, y alrededores.

En general, los datos provistos por las instituciones mencionadas ya han sido consistidos, se han detectado outliers y se han corregido. Sin embargo, igualmente fueron cuidadosamente controlados de forma de asegurar que no tuvieran errores residuales. Los datos faltantes en cada estación

Tesis - Juan Ignacio Mazza

meteorológica representan menos del 18% del total y fueron completados con la mediana de la precipitación mensual de la estación en cuestión.

El formato Shapefile (SHP) es un formato de archivo informático propietario de datos espaciales desarrollado por la compañía ESRI, quien crea y comercializa software para Sistemas de Información Geográfica. Se utilizaron archivos shapefile para los polígonos de Argentina y de Gran chaco a fin de localizar sobre ellos las estaciones meteorológicas.

CAPÍTULO III

MÉTODOS Y RESULTADOS

3.1. Regionalización del área de estudio

Como se mostró en el capítulo anterior, se cuenta con los registros acumulados de precipitación mensual para 34 estaciones meteorológicas en el área de Gran Chaco, Argentina. Debido a que los forzantes climáticos actúan sobre regiones espacialmente extensas, resulta de interés agruparlas de algún modo, intentando que, estaciones meteorológicas con regímenes similares de lluvia pertenezcan al mismo grupo. Para esto se utilizaron redes neuronales de Kohonen.

Las redes neuronales de Kohonen, o mapas auto-organizativos (SOM, por sus siglas en inglés) son redes neuronales no supervisadas capaces de determinar grupos de estaciones (clusters) que registran comportamientos de lluvia similares.

Se realizaron varias pruebas utilizando:

- Una red de 4 neuronas (grilla de 2x2) y las mediciones de los 12 meses del año
- Una red de 9 neuronas (grilla de 3x3) y las mediciones de los 12 meses del año
- Una red de 4 neuronas (grilla de 2x2) y las mediciones de las 4 estaciones del año
- Una red de 9 neuronas (grilla de 3x3) y las mediciones de las 4 estaciones del año

El método se aplicó para las mediciones de precipitación acumulada mensual en los 12 meses del año. La cantidad de neuronas que se utilizan resulta un elemento importante para la clasificación. Una red de Kohonen de 9 neuronas genera grupos demasiado reducidos (en algunos casos de sólo dos estaciones) por lo que finalmente se decidió proceder con una red de 4 neuronas (grilla de 2x2). Debido a la utilización de una red de 4 neuronas, se obtuvieron 4 grupos que muestran comportamientos similares en cuanto a precipitación. La figura 10 muestra dichos grupos junto con los valores de precipitación mensual media.

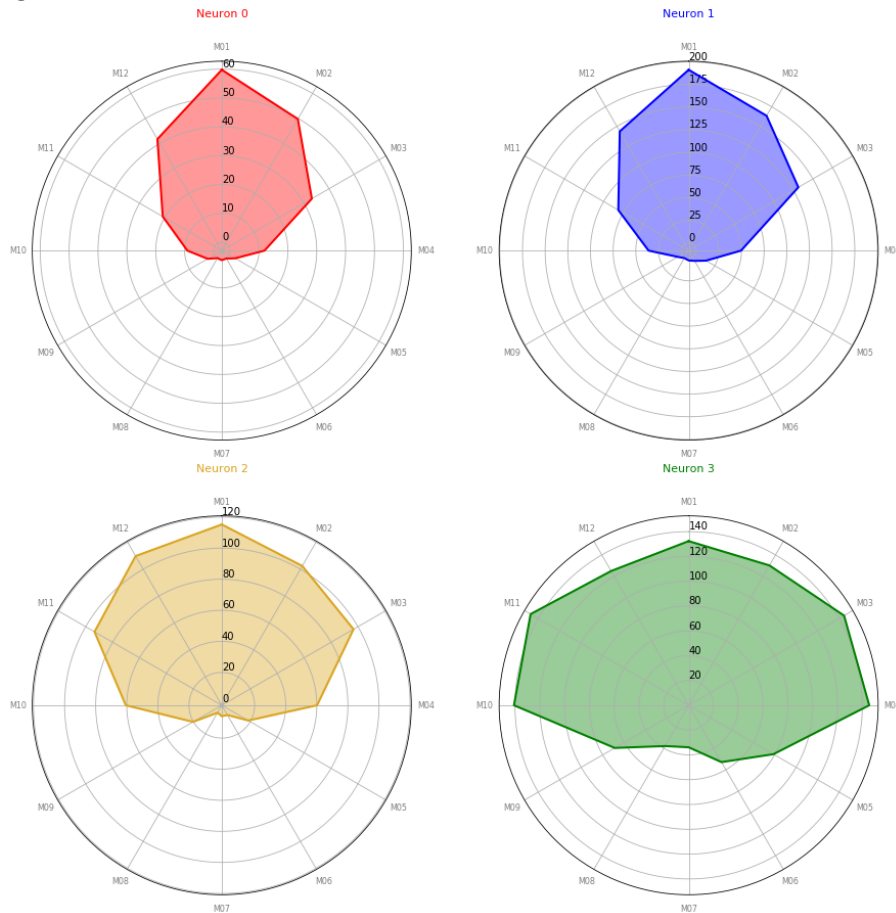


Figura 10. Neuronas de la red de Kohonen en la agrupación de estaciones meteorológicas

Algunas conclusiones que podemos sacar de la figura anterior son:

- La neurona 0 (en rojo) agrupa a las estaciones con lluvia baja (aproximadamente 40mm) en los meses de verano (finales de diciembre, enero, febrero y parte marzo) y lluvias prácticamente nulas el resto del año.
- La neurona 1 (en azul) es similar a la anterior, agrupa las estaciones con lluvia alta (aproximadamente 175mm) en los mismos meses de verano y muy baja precipitación el resto del año.
- La neurona 2 (en amarillo) agrupa estaciones con lluvia máxima en los meses de verano (aproximadamente 100mm), sin embargo, las estaciones que pertenecen a este grupo registran lluvias moderadas desde la primavera (aproximadamente 60mm en octubre, 90mm en noviembre y 110mm en diciembre). Finalmente, cabe mencionar que la época invernal registra menor precipitación que el semestre estival.

Tesis - Juan Ignacio Mazza

- La neurona 3 agrupa estaciones que registran altas lluvias durante todo el año aunque se observa también menor cantidad en invierno (del orden de 60mm) que en verano (del orden de 120mm).
- Los 4 grupos tienen en común la reducción en la precipitación medida entre los meses de mayo y septiembre, es decir que en toda la región se visualiza una onda anual de precipitación con máximos en la estación estival.

Se muestran a continuación (Figura 11) las estaciones meteorológicas pertenecientes a cada grupo dibujadas sobre un mapa de la República Argentina. Algunas consideraciones a tener en cuenta son:

- Los colores utilizados se corresponden con los colores del grafico anterior.
- El diámetro del marcador se corresponde con la variable [altitud], siendo el punto de mayor diámetro el que representa la estación de La Quiaca.

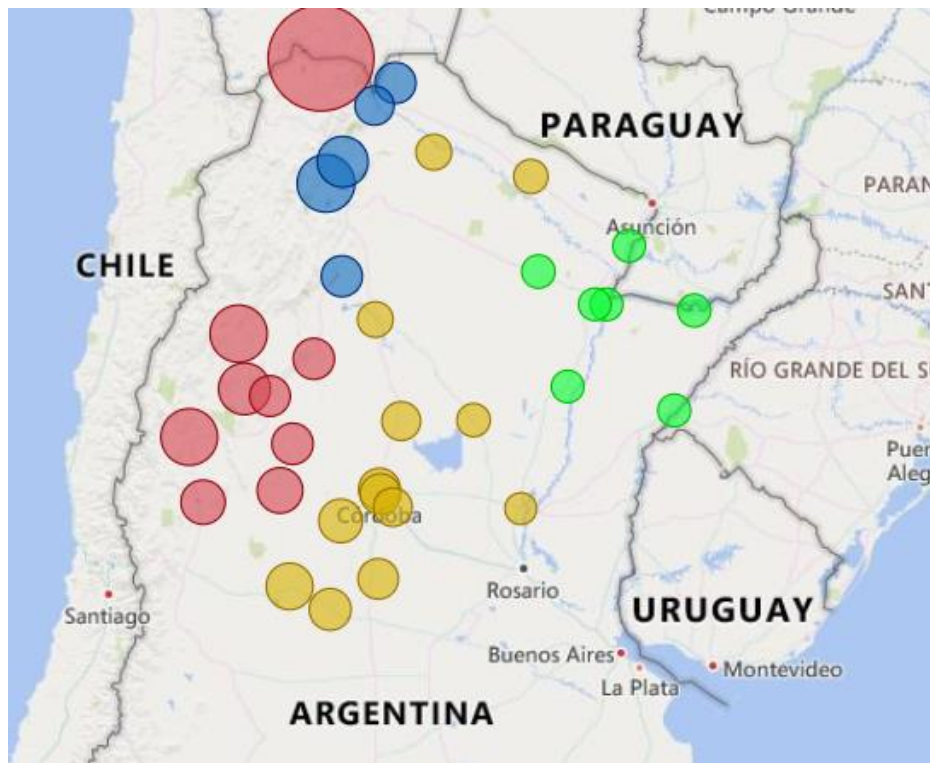


Figura 11. Estaciones meteorológicas agrupadas dibujadas sobre un mapa de la Argentina

El agrupamiento es coherente con los regímenes de precipitación observados, mostrando hacia el oeste la región más seca y hacia el este la más húmeda. La mayor precipitación de verano se produce por el ingreso de aire húmedo desde la selva brasilera en dicha época del año [19].

Por otra parte, podemos observar que la estación meteorológica de La Quiaca, pertenece al grupo 0 (rojo) y no al grupo 1 (azul) como sus estaciones vecinas. Esto se debe a que, debido a la altitud de La Quiaca, su comportamiento en cuanto a lluvias es más parecido al de las estaciones cercanas a la cordillera que al de las estaciones geográficamente más cercanas.

El listado completo de estaciones meteorológicas junto con su grupo asignado se encuentra disponible en el ANEXO 1.

Una vez obtenidos los grupos, resulta de interés evaluar si efectivamente los valores mensuales de lluvia acumulada que obtuvo cada neurona de la red de Kohonen se corresponden con las mediciones reales de las estaciones pertenecientes a dicho grupo. Las figuras 12, 13, 14 y 15 muestran la variabilidad interanual de la precipitación mensual en algunas estaciones representativas de cada grupo.

Cabe destacar que, con el objetivo de obtener visualizaciones más claras, todas las series temporales fueron “suavizadas” utilizando la técnica **Promedios Móviles** con parámetro de 5 términos (**MA = 5**).

El listado completo de gráficos se encuentra disponible en el ANEXO 2.

Grupo 0

Entre las estaciones que conforman el grupo 0 se encuentran **Chepes** y **La Quiaca**. En ambas podemos notar como la precipitación de los meses de verano (enero, febrero y la mayor parte de marzo, en azul) fluctúan en su mayoría entre los 40mm y 100mm y entre abril y septiembre (representados en rojo y verde) las lluvias son muy bajas, menores a 20mm.

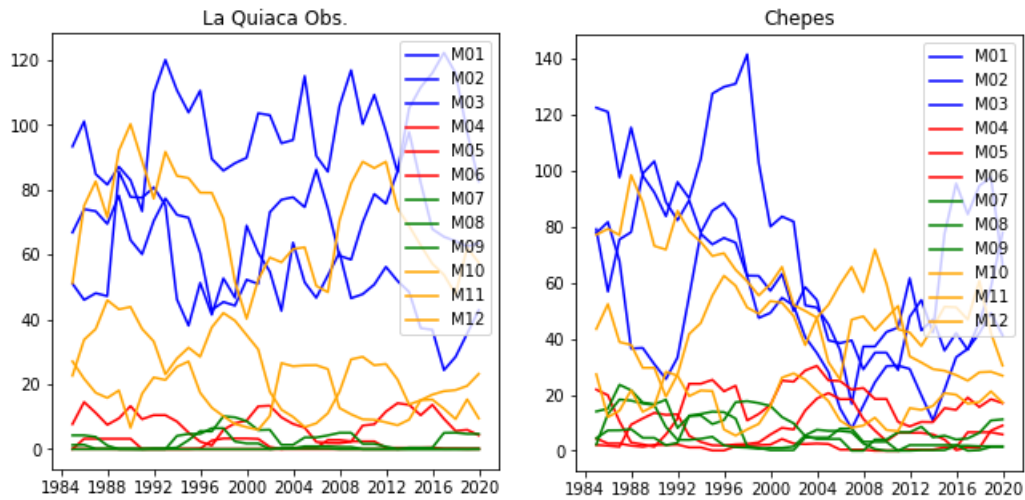


Figura 12. Series temporales de precipitación mensual acumulada para dos estaciones del grupo 0

Grupo 1

Dentro del grupo 1 encontramos a **Tartagal** y **Jujuy**, tal como quedara expresado en la neurona del SOM que representa a este grupo, ambas locaciones muestran lluvias alrededor de los 175mm para los meses de verano (en azul) y lluvias muy bajas para los meses de abril a septiembre. El mes de diciembre muestra un comportamiento de transición, ya que muchos años la precipitación ocurrida ese mes es alta, respondiendo al régimen de verano. Cabe destacar que Tucumán ha tenido años particularmente lluviosos con picos muy anómalos de casi 300mm.

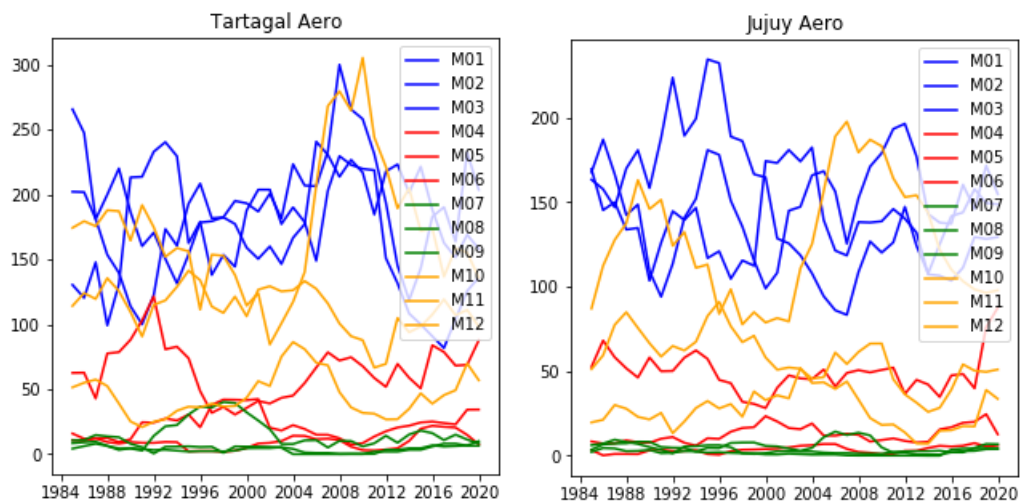


Figura 13. Series temporales de precipitación mensual acumulada para dos estaciones del grupo 1

Grupo 2

A diferencia de los grupos anteriores, en el grupo 2 no existe una diferencia tan marcada entre la primavera (octubre, noviembre y diciembre) en amarillo, y los que conforman el verano (enero, febrero, marzo) en azul. Como podemos observar en la figura 14, esto se corrobora para las estaciones de **Ceres** y **Córdoba**. Además, tal como se observa en la neurona 2 de la red de Kohonen, los restantes meses del año (abril a septiembre) presentan lluvias bajas.

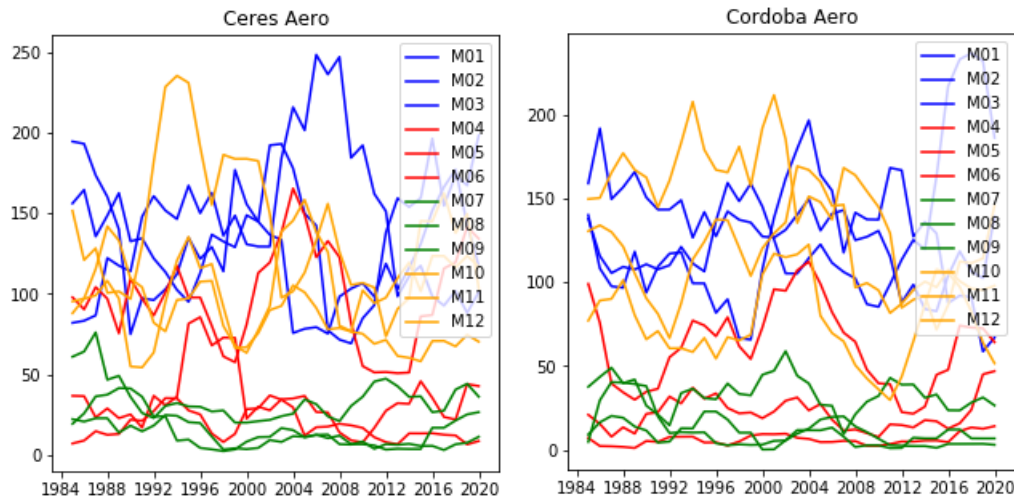


Figura 14. Series temporales de precipitación mensual acumulada para dos estaciones del grupo 2

Grupo 3

Como puede verse en la neurona 3 de la red de Kohonen, el último grupo es el que presenta la mayor similitud entre los meses de primavera y los de verano. Esto también se observa para las estaciones de **Corrientes** y **Formosa** donde las líneas que corresponden a la primavera (octubre, noviembre y diciembre) en amarillo se asemejan con las líneas que corresponden a los meses de verano (enero, febrero, marzo) en azul. Además, al igual que en los grupos anteriores, se observan menores precipitaciones entre abril y septiembre.

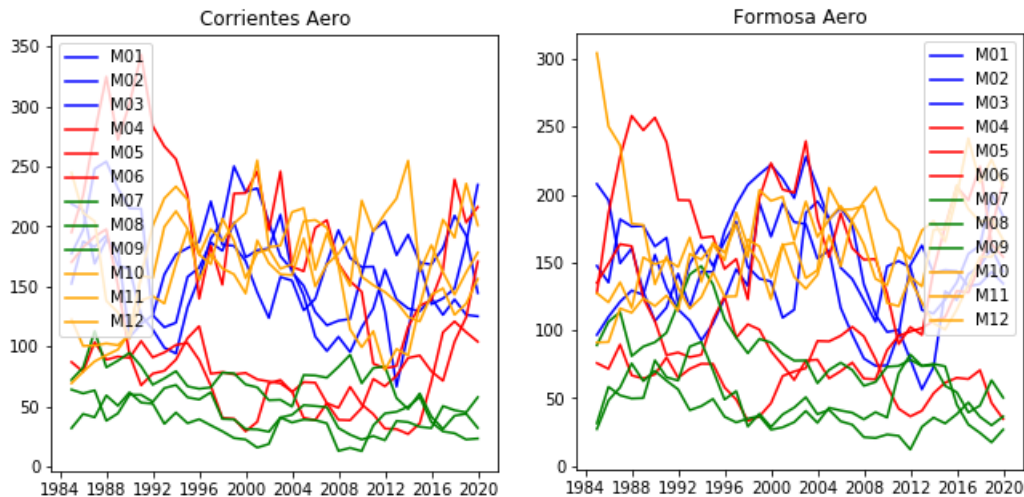


Figura 15. Series temporales de precipitación mensual acumulada para dos estaciones del grupo 3

Una vez definidos los grupos y las estaciones que los componen, se procedió a calcular la serie mediana de los mismos, por ejemplo, el valor de enero de 1980 para el grupo 0 (G0) es la mediana de todos los valores de enero de 1980 de las estaciones que componen dicho grupo.

En los 40 años analizados, varias estaciones han reportado valores excepcionalmente anómalos para algún mes, de aquí surge la decisión de utilizar la mediana como medida más representativa en lugar de la media, ya que esta no es una medida sensible a valores extremos.

El listado completo de series de precipitación medianas mensuales acumuladas se encuentra disponible en el ANEXO 3.

3.2. Determinación de Predictores Análisis de Correlación

Una vez definidos los grupos y calculadas sus series medianas, se procedió a calcular la correlación entre ellas y las variables meteorológicas y oceánicas globales. Físicamente, se busca establecer aquellos posibles forzantes climáticos que afectan a la precipitación en cada región. Para ello fue necesario establecer una causa física que sustente la correlación en el caso de que la misma fuera significativa. Este proceso constó de una serie de pasos que se detallan a continuación.

Las variables meteorológicas globales fueron procesadas y el conjunto de datos resultantes fueron matrices de 73 x 144 que abarcan la superficie de la tierra y en cada punto de reticulado cada variable es una serie temporal de 480 términos. Se tienen entonces para cada una de las 7 variables predictoras elegidas (HGT200, HGT500, HGT1000, TCW, SST, U850 y V850) 10.512 series temporales (una por cada posición de la grilla de 73 x 144).

Para la serie temporal de cada variable en cada punto de retícula, es posible calcular la correlación con la serie mediana de precipitación de un determinado mes de cada uno de los cuatro grupos obtenidos, utilizando la correlación de Pearson.

Es importante destacar que, dado que la intención de este trabajo es pronosticar lluvias a mediano plazo, las correlaciones que se buscan tienen un mes de desfase. Es decir, nos interesa observar una relación relevante (correlación significativa) entre el comportamiento de la variable predictora en un mes determinado y la precipitación ocurrida el siguiente mes. Por ejemplo, el comportamiento de los predictores en marzo y la precipitación de abril.

La figura 16 muestra a modo de ejemplo, el campo espacial de correlaciones obtenidas entre las series temporales de enero en cada punto de retículo [M01] de la variable predictora HGT500 y la serie temporal de medianas de precipitación en febrero [M02] para el grupo 0 [G0]. Cada pixel de la imagen representa la correlación entre la serie temporal de la variable para ese par de [Latitud, Longitud] y la serie temporal del grupo.

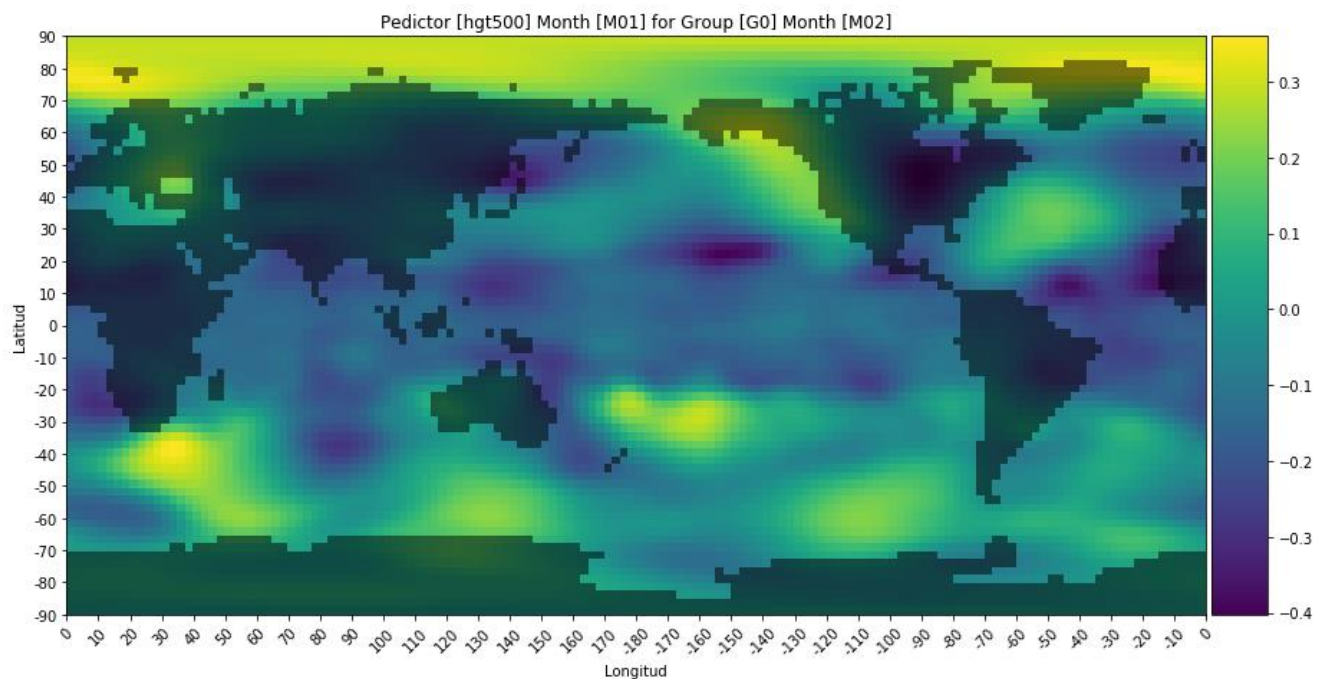


Figura 16. Mapa de correlación entre las series de enero de HGT500 y febrero del grupo 0

3.2.2. Definición de predictores

Una vez finalizado el proceso anterior, se habrán calculado más de 294 mil correlaciones de Pearson, sin embargo, no todas ellas serán relevantes para predecir la precipitación en el Gran Chaco Argentino. La primera cuestión que surge entonces, es establecer un umbral a partir del cual la correlación se considerará significativa para este análisis. La significancia se calculó utilizando el test Normal con un 95% de confianza. De esta forma resultaron significativas las correlaciones donde el coeficiente de Pearson superó en módulo el valor dado por la siguiente expresión:

$$\text{Limite} = \pm \frac{1.96}{\sqrt{n-2}}$$

Fórmula 1. Intervalo de confianza utilizado en el cálculo de una correlación significativa con 95% de confianza.

Donde n es el número de datos de la serie. De esta fórmula surgen dos valores que forman un intervalo [-0.3179; +0.3170], cualquier valor de correlación de Pearson que no esté contenido en dicho intervalo, es decir, valores menores que -0.3179 o mayores que +0.3179 serán considerados significativos con el 95% de confianza, por lo tanto, los valores que se encuentren dentro del intervalo serán considerados no significativos y descartados.

Una vez aplicado el filtro anterior, habremos descartado aquellas series de tiempo cuya correlación resultó no significativa. Sin embargo, esto no es suficiente ya que cada pixel del mapa, es decir, representan un área relativamente pequeña de la Tierra: los puntos distan entre si 2.5° (unos 277 km, aproximadamente). Los forzantes climáticos refieren a regiones medianamente extensas para asegurar que la señal no es casual sino que responde a un proceso físico asociado. Es por ello que, para definir predictores, se busca un área donde haya varios pixeles que muestran la misma señal.

Surge entonces la necesidad de detectar áreas con alta significancia. En este trabajo se consideran aquellas que contienen al menos 15 pixeles contiguos (4155km x 4155km como mínimo) para los cuales su serie de tiempo asociada tuvo una correlación significativa con la serie de precipitaciones.

La figura 17 muestra a modo de ejemplo las áreas de alta significancia para las series temporales del mes de octubre de la variable que mide la temperatura de la superficie del mar (SST, por sus siglas en inglés) con respecto a la serie de precipitación mediana del mes de noviembre del grupo 1 (G1). Se detalla para cada área la media de todos los valores de correlación de los pixeles que la componen.

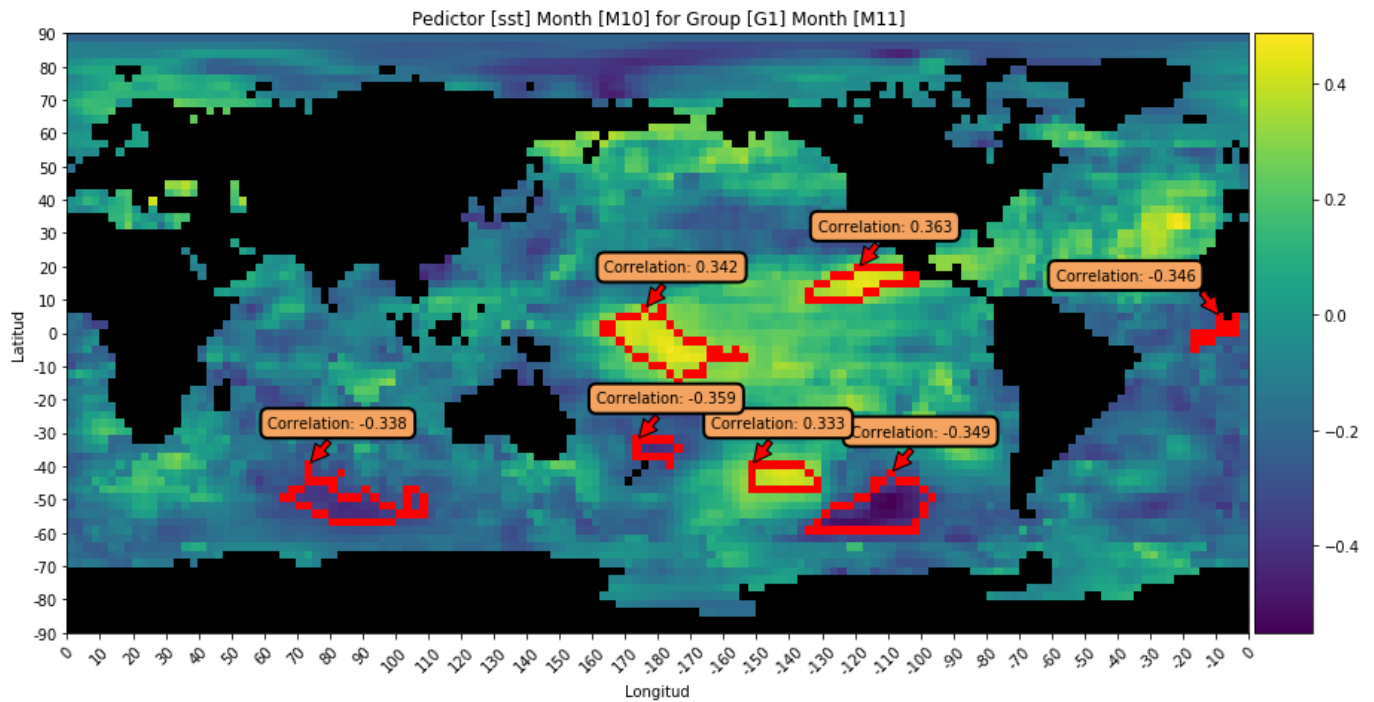


Figura 17. Ejemplo de áreas de correlación significativa entre la variable SST en octubre y la precipitación de noviembre en el grupo 1

Una vez finalizado el proceso anterior se habrán generado 336 mapas con múltiples áreas significativas como el que se mostró en la figura 9. Sin embargo, no todas estas áreas significativas son realmente relevantes a efectos de predecir la precipitación en Gran Chaco. Muchas de ellas podrían haber resultado significativas sin que las respalde un proceso físico explicable. Debido a que alta correlación no siempre representa causalidad, es importante poder explicar físicamente la forma en que cada predictor influencia la precipitación. Por ejemplo, se sabe que en Argentina Subtropical en muchas circunstancias domina el viento con componente norte o este, es decir viento proveniente de la selva brasilera o del océano Atlántico Sur respectivamente. Estos vientos transportan aire húmedo que favorece la precipitación. Por lo tanto, es de esperar que variables predictoras situadas en estos lugares sean importantes de considerar. También hay que considerar los mecanismos más complejos llamados “teleconexiones” que se detallaron en la introducción y que vinculan anomalías en áreas distantes con la de precipitación en Argentina. Probablemente el más conocido es el fenómeno “El Niño” que relaciona anomalías cálidas en el océano Pacífico tropical central con lluvias en Argentina subtropical, pero no es el único. Todos estos elementos fueron tenidos en cuenta a la hora de diseñar las máscaras que se detallan a continuación.

Para cada una de las 7 variables de estudio, se buscó que el área de correlación significativa sucediera en un lugar de la Tierra donde pudiera ser explicada mediante un fenómeno físico o

Tesis - Juan Ignacio Mazza

atmosférico. Para ello se utilizaron máscaras de filtrado, imágenes definidas por los usuarios (meteorólogos con conocimiento de dominio) que se superponen con los mapas antes mostrados y descartan todo lo que quede por fuera de las áreas definidas.

Las máscaras pueden ser definidas mediante un archivo .csv utilizando el siguiente formato, todas las máscaras definen áreas cuadradas o rectangulares, por lo que necesitan 4 valores:

<NombreDeLaVariable>:<LongLimite1>,<LongLimite2>,<LatLimite1>,<LatLimite2>

Por ejemplo:

hgt200:0,-90,-180,180

Resulta en la siguiente máscara:

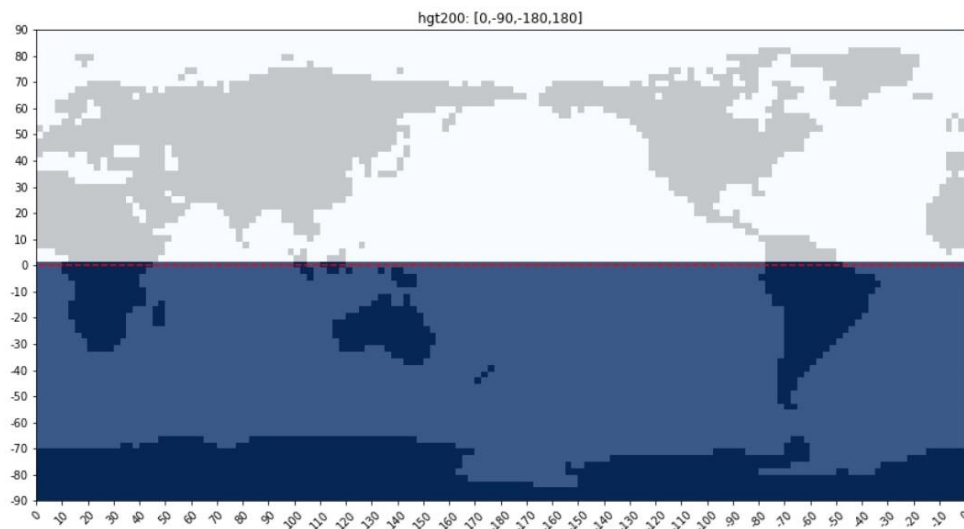


Figura 18. Máscara de filtrado para la variable HGT200

También es posible para una misma variable, definir más de una máscara separando los límites con el caracter “|”. El siguiente ejemplo muestra dos mascarar definidas para la variable u850:

u850:-22,-34,-68,-58|-30,-70,0,150

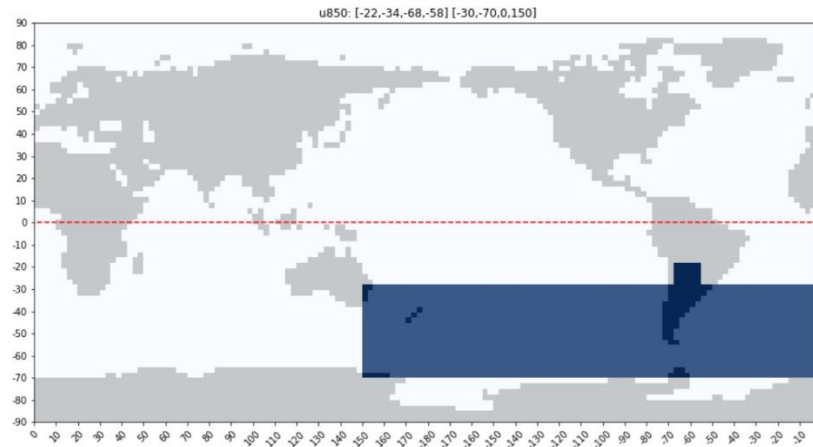


Figura 19. Máscara de filtrado para la variable U850

El listado completo de máscaras de filtrado para cada variable junto con su definición en términos de latitudes y longitudes se encuentra disponible en el ANEXO 4. Algunas observaciones merecen ser detalladas para comprender la definición de estas máscaras. Las variables relacionadas con la circulación (Hgt200, Hgt500, Hgt1000) se considerarán sólo en el hemisferio sur pues no hay intercambio de masa entre hemisferios. La variable SST se considera en zonas tropicales (20°N- 60°S) porque muchos fenómenos de teleconexión se relacionan con índices definidos en áreas que abarcan el hemisferio norte, como por ejemplo el fenómeno ENOS que se define por las anomalías de SST en el Pacífico tropical central, abarcando parte en ambos hemisferios. Por otro lado, los datos de reanálisis trabajan con información satelital que falla en detectar la SST en latitudes altas, por lo cual la máscara utilizada elimina estas latitudes. En el caso de la variable U850, se considera que el flujo de los oestes en la zona del Pacífico y Atlántico en latitudes medias es relevante pues representa la trayectoria de las ondas de Rossby asociadas a sistemas precipitantes y por ello la máscara se ha definido como (30°S-70°S; 150°E-0). Para las variables V850 y Tcw la misma se definió en la región de la zona estudiada del Gran Chaco (0-40°S; 65°O-35°O) debido a que son importantes los vientos locales provenientes de la selva brasilera y del anticiclón del océano Atlántico sur, que canalizan humedad hacia el sur en mayor o menor medida, causando inestabilidad y por lo tanto precipitación.

Una vez generados los mapas con las áreas de correlación significativa y las máscaras de filtrado definidas en el punto anterior, se procedió a la aplicación de las mismas. El proceso es relativamente sencillo, la máscara de filtrado se superpone con el mapa generado y sólo las áreas de alta significancia que quedan fuera de la máscara son conservadas. Cabe destacar que, si la máscara contiene parcialmente un área, solo la parte que cae dentro de la máscara es eliminada.

La figura 20 muestra, por ejemplo, la aplicación de la máscara de la variable HGT200, la altura geopotencial sólo se considera si ocurre en el hemisferio sur, por lo que cualquier área del hemisferio norte es eliminada.

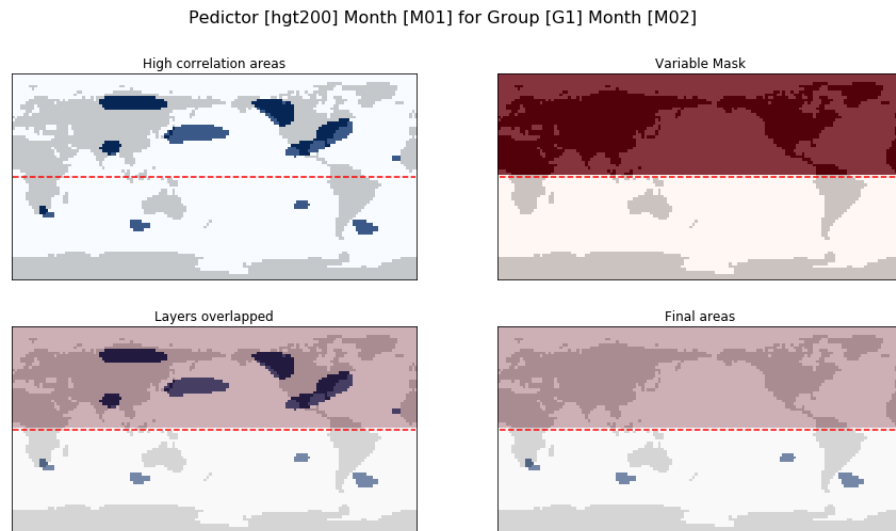


Figura 20. Ejemplo de aplicación de una máscara de filtrado para la variable predictora HGT200.

El siguiente ejemplo (figura 21), muestra la aplicación de la variable de viento U850, nótese como las áreas que caen parcialmente dentro de la máscara son recortadas.

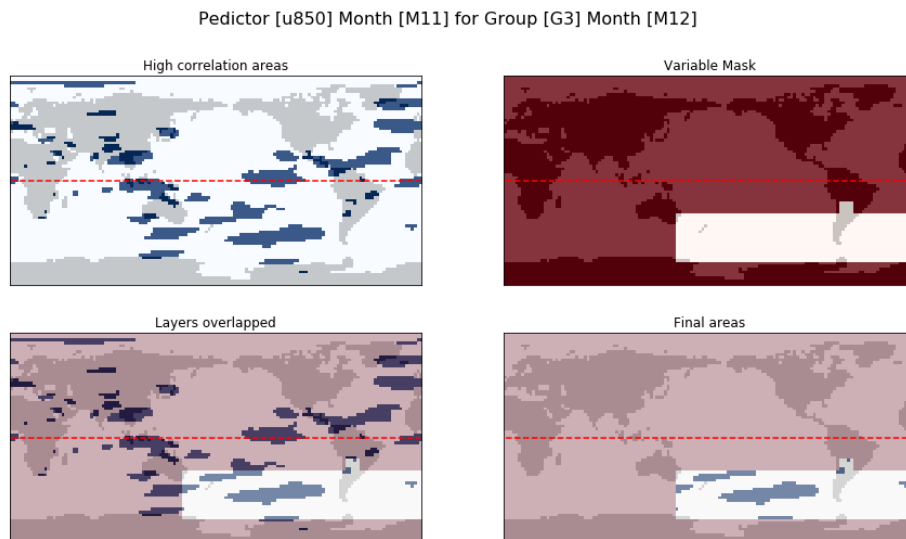


Figura 21. Ejemplo de aplicación de una máscara de filtrado con recorte de área para la variable U850.

Finalmente, una vez aplicado el proceso de filtrado, las áreas de interés para el análisis cumplen dos requisitos fundamentales:

Tesis - Juan Ignacio Mazza

- Cumplen la condición de definirse con un área suficientemente grande como para asegurar que no es una correlación casual
- Suceden en algún lugar de la Tierra que puede ser explicado mediante un fenómeno meteorológico conocido

El paso final consiste en obtener los predictores con los que se entrenarán los modelos de machine learning a partir de ellas.

Recordemos que cada área está compuesta por píxeles, y cada uno de estos píxeles representa una serie temporal mensual de mediciones de alguna de las variables en un determinado par de latitud y longitud. El objetivo es entonces, obtener a partir de estas múltiples series de los píxeles, una única serie que represente al área, para ello simplemente se calculó la serie media de todas las series que la componen.

Como resultado final de este proceso, se obtuvieron 12 archivos de predictores por cada uno de los cuatro grupos (48 archivos en total), además, dada la naturaleza misma de este enfoque, es perfectamente posible que cada uno de estos archivos contenga una cantidad distinta de predictores dependiendo de cuantas áreas de correlación se encontraron y su ubicación a la hora de filtrarlas.

La figura 22 muestra algunos predictores para las lluvias de enero del grupo 1 (G1). La nomenclatura de los encabezados sigue la siguiente regla:

NombreVariable_NumeroMesDeLaVariable_IdentificadorDeArea

Por ejemplo, para la primera columna el encabezado queda:

NombreVariable: hgt1000 (la variable es la altura geopotencial en 1000Hpa, capas bajas)

NumeroMesDeLaVariable: M12 (el mes en que se calcula la variable es diciembre, recordemos que se predice con un mes de desfase, por lo que las precipitaciones de enero se predicen utilizando las variables de diciembre del año anterior)

IdentificadorDeArea: La letra A seguida de un número identificatorio del área de la variable predictora, para garantizar unicidad en los nombres de las columnas.

Tesis - Juan Ignacio Mazza

hgt1000_M12_A00	hgt1000_M12_A01	hgt1000_M12_A02	hgt1000_M12_A03	hgt500_M12_A03	sst_M12_A00	sst_M12_A01	tcw_M12_A00	u850_M12_A00
108.401344	132.9937966	6.758736472	-74.05627003	5142.928178	29.73927155	-18.1940808	39.98301809	0.217998047
102.0709675	102.2196029	47.80208318	-89.50885265	5119.85655	29.46089244	-21.67662953	38.60933618	4.803198853
109.0981181	141.8188584	123.6878351	-142.8826096	5073.401223	29.00339904	-17.0726043	37.18762948	6.323599243
98.85752684	78.27295291	32.24865567	-38.65316514	5209.178426	30.10098848	-17.21086531	37.94785533	0.059196777
101.3293011	119.029156	50.88709715	-107.4838709	5094.028722	29.30776296	-17.21443078	37.31753383	3.574797363
96.3475804	133.7375931	16.24025526	-38.83264611	5193.001526	29.44334345	-22.15902623	34.09489811	2.833599243
103.4215054	103.1284114	61.74966385	-131.4472473	5057.989046	29.01686816	-22.50242925	32.82990658	7.834797363
104.7045697	124.4658808	76.72278202	-133.4557364	5086.922786	29.40793242	-19.48852219	36.38905133	4.787196655
110.4494624	114.9342429	34.91095448	-102.3405286	5109.720638	28.79609361	-18.41365859	35.11203093	2.893998413
106.4456986	133.3728281	75.87063141	-121.3206404	5068.458394	28.72691641	-18.02720428	35.54950647	5.859998169
100.7153226	85.32506224	73.8965054	-97.40286218	5150.992845	29.57622213	-19.35125214	41.9772478	3.838797607
102.1088707	87.12034695	54.63306447	-103.2165897	5141.815473	29.61078725	-18.06394649	34.09969554	0.69119873
104.6311829	84.82195979	37.5987902	-40.57070099	5218.831972	29.31981988	-21.24122891	35.32332129	4.23159729
108.4744625	104.352978	51.741599	-111.378608	5082.989155	29.11368999	-19.05560727	35.29476289	9.028397217
107.4508066	106.0310172	54.38340086	-147.2466653	5043.439436	29.50405221	-20.00796322	38.32047339	8.617597046
111.9489251	92.98511182	62.58366884	-151.1794808	5045.587928	29.79097257	-21.47661036	39.39806344	3.334797363
110.7456989	155.5471466	74.85618235	-129.7395097	5062.651998	29.12119083	-17.99161357	37.866206	3.650797729
102.8908604	117.7258065	60.25201614	-86.36987616	5140.817105	29.76119914	-20.34721563	38.63047454	4.30119873
107.7602154	117.3368485	60.77788978	-75.78898855	5142.672176	29.98792968	-19.4464538	38.02219885	1.702799683
106.9836023	107.014268	87.52452922	-137.7060395	5066.727008	29.55601897	-17.96584367	36.06481956	4.856797485
103.9500002	110.0955337	56.81182762	-156.2107683	5014.575289	29.72571235	-18.46972406	36.47785411	6.330797119

Figura 22. Ejemplo de predictores para la precipitación de enero del grupo 1

Se muestra a continuación (figura 23) para el mismo grupo y mes, un correlograma de las variables predictoras y la precipitación del grupo 1 en enero (denotada como 'Rain'). El mismo muestra la correlación entre cada par de variables predictoras y en la última columna la correlación de cada predictor con la precipitación. Se puede observar que es una matriz simétrica. Además, es importante notar que la correlación entre pares de predictores es alta en algunos casos (verdes y rojos oscuros), superando el valor crítico para ser consideradas significativas con el 95% de confianza. Esto indica que físicamente esos predictores no son independientes entre sí. Este hecho es una situación común en meteorología pues las variables están muy relacionadas unas con otras y es difícil aislar fenómenos independientes. Desde el punto de vista estadístico y de selección de predictores esto puede resultar un inconveniente difícil de resolver. Cuanto mas independientes sean los predictores mejor se afronta el problema de la multicolinealidad.

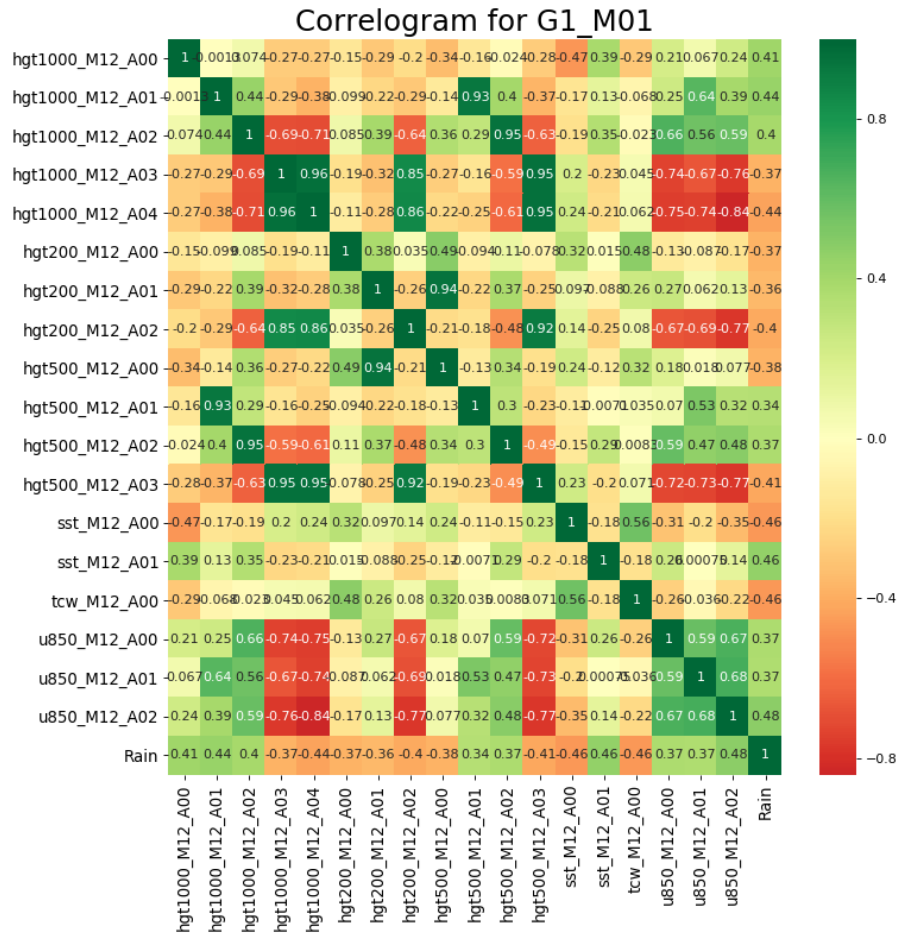


Figura 23. Correlograma entre las variables predictoras y la clase para enero del grupo 1 [G1]

3.3. Diseño de modelos estadísticos

Todos los modelos generados con los distintos métodos de machine learning que se detallan a continuación (redes neuronales, XGBoost, regresión de soporte vectorial y el ensamble de los mismos) fueron entrenados de la misma manera. Se utilizaron los años 1980-2008 como periodo de entrenamiento y los años 2009-2019 como periodo de verificación.

Si bien el resultado de los modelos predictores es la cantidad de lluvia pronosticada en milímetros (modelos de regresión), es habitual dar los pronósticos en términos categóricos, por ejemplo: Cantidad de lluvia esperada normal, subnormal o supranormal.

Por este motivo los resultados obtenidos son siempre discretizados en tres categorías. Para cada par [Grupo-Mes] se calculan los terciles utilizando los datos de entrenamiento. Finalmente se utilizan estos terciles para determinar la categoría de la lluvia pronosticada:

- **Primer tercil:** Lluvia subnormal para el Grupo-Mes
- **Segundo tercil:** Lluvia normal para el Grupo-Mes
- **Tercer tercil:** Lluvia supranormal para el Grupo-Mes

3.3.1. Redes neuronales artificiales

Las redes neuronales artificiales, son modelos matemáticos y computacionales que consisten en un conjunto de neuronas interconectadas entre sí para transmitirse señales. La figura 24 muestra un ejemplo básico de una arquitectura de red neuronal.

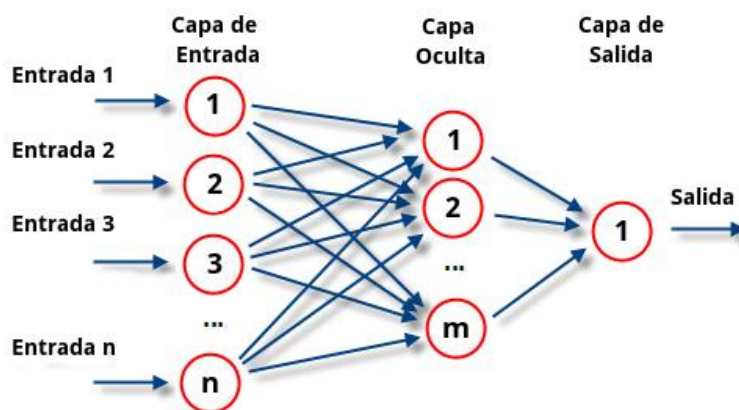


Figura 24. Ejemplo de red neuronal

Como podemos observar, las neuronas se encuentran organizadas en capas pudiendo distinguirse las siguientes:

- Una **capa de entrada**, que recibe los inputs, esta capa siempre tiene tantas neuronas como valores de entrada tenga la red
- Una o más capas intermedias, conocidas como **capas ocultas**, para las cuales el número óptimo de neuronas es desconocido
- Una **capa de salida** que genera el output de la red, la cantidad de neuronas en esta capa depende de la salida esperada de la misma, para este trabajo, y dada la naturaleza regresora de la red, la capa de salida consta de una única neurona que devuelve el valor final de lluvia predicho

Las neuronas de cada capa se conectan con las neuronas de la capa siguiente. La información de entrada fluye a través de cada capa mientras atraviesa la red neuronal, donde se somete a diversas operaciones (cálculo de entrada neta de las neuronas ponderando las entradas con pesos y la aplicación de funciones de activación) produciendo uno o más valores de salida.

El objetivo del entrenamiento de estos modelos es hallar un conjunto de pesos W , tal que W minimiza la función de pérdida. Para este trabajo se ha decidido trabajar con el error absoluto medio (MAE por sus siglas en inglés) de la red. Existen varios métodos de optimización para minimizar dicha función, para este trabajo se han utilizado los métodos nadam y adadelta.

Finalmente, con el objetivo de prevenir el sobreajuste, se utilizó la técnica de regularización "Dropout". Esta técnica funciona de la siguiente manera: por cada nueva entrada a la red en fase de entrenamiento, se desactivan aleatoriamente un porcentaje de las neuronas en cada capa oculta acorde a una probabilidad de descarte previamente definida. Dicha probabilidad puede ser igual para toda la red, o distinta en cada capa. Lo que se consigue con esto es que ninguna neurona memorice parte de la entrada; que es precisamente lo que sucede cuando tenemos sobreajuste.

A continuación se detalla la aplicación de las redes neuronales en el caso del pronóstico de la precipitación en el Gran Chaco Argentino. Una vez obtenidas las variables predictoras para la precipitación de cada mes de cada grupo, se entrenaron diversas arquitecturas de redes neuronales para generar dichos pronósticos.

Cabe destacar que, como se explicó anteriormente, existe una gran variedad de parámetros que hacen a una arquitectura de red neuronal, por ejemplo: cantidad de capas ocultas, cantidad de neuronas por capa oculta, función de activación de cada capa, tipo de optimizador, cantidad de épocas, dropout, etc. Esto hace que el número de redes posibles sea tan grande que es necesario elegir con cuáles trabajar.

Para este trabajo, se probaron diversas arquitecturas de redes neuronales variando los siguientes parámetros:

- Cantidad de capas ocultas: 2; 6
- Neuronas por capa oculta: 64; 128
- Funciones de activación: lineal, ReLU
- Optimizadores: nadam, adadelta
- Dropout: 0; 0,1; 0,25

Es decir, que, para cada mes de cada grupo, se entrenaron 48 redes neuronales con el objetivo de determinar la arquitectura que mejor captaba la estructura de los datos.

Se hicieron además pruebas con tamaño de batch igual a 1 y con optimización por “Descenso estocástico de gradiente” (o SGD por sus siglas en inglés) con y sin momentum de Nesterov. Ambos parámetros demostraron nunca dar buenos resultados (las redes con menor error nunca los incluían) por lo que fueron eliminados de los hiperparámetros a explorar con el objetivo de reducir la cantidad de combinaciones de parámetros posibles y por lo tanto de redes a probar.

3.3.2. Regresión de Soporte Vectorial

Las Máquina de Soporte Vectorial (SVM por sus siglas en inglés) es un algoritmo de clasificación propuesto en 1995 por Vladimir Vapnik [48]. El concepto fundamental de la máquina de soporte vectorial es una idea extremadamente sofisticada con una implementación realmente sencilla: los vectores de entrada se mapean de forma no lineal a un espacio de mayor dimensionalidad conocido como “feature space”. En este espacio, se construye una superficie de separación, capaz de discriminar entre clases, que es utilizada para etiquetar las observaciones (Figura 25).

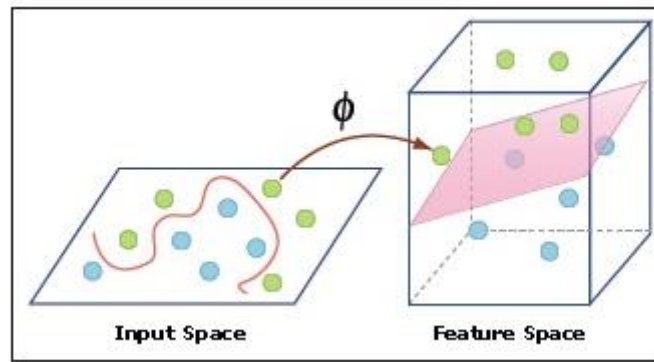


Figura 25. Proyección de los vectores de entrada a un espacio de mayor dimensionalidad.

Este mapeo a un espacio de mayor dimensionalidad es realizado por una función denominada Kernel, si bien esta función puede ser cualquiera definida por el usuario, existen algunas variantes predefinidas como ser:

- Kernel radial
- Kernel lineal
- Kernel polinómico: Requiere definir el grado del polinomio y, opcionalmente, el término independiente (coef)
- Kernel sigmoide: Opcionalmente puede incluir el término independiente (coef)

Una vez aplicado el Kernel y con los datos ya proyectados a un espacio de mayor dimensionalidad, la máquina de soporte vectorial intentará buscar un hiperplano (o superficie de separación) que cumpla las siguientes dos condiciones:

- El hiperplano es de margen máximo (está lo más alejado posible de las clases que separa)
- El hiperplano clasifica correctamente la mayor cantidad posible de instancias

Sin embargo, no siempre es posible cumplir ambas condiciones simultáneamente, es por esto que existe un parámetro de regularización llamado 'C' que permite 'sacrificar' en parte la primera condición para maximizar la cantidad de instancias correctamente clasificadas. En la figura 26 se muestra un esquema de cómo un mayor valor de 'C' influencia el algoritmo para obtener un hiperplano que no es de margen máximo, pero clasifica correctamente todas las observaciones:

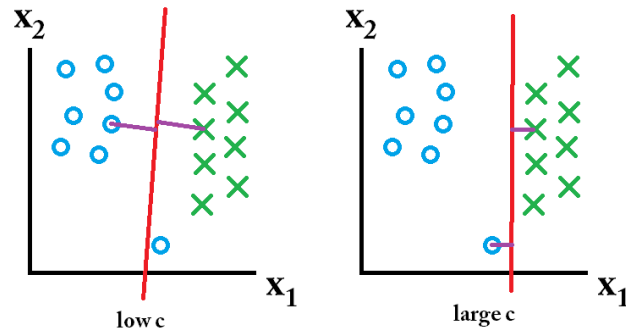


Figura 26. Influencia del parámetro 'C' en el hiperplano de separación de una máquina de soporte vectorial

El propio Vladimir Vapnik propuso en 1996 una versión de la máquina de soporte vectorial que podía ser utilizado para regresión y a la que llamo Regresión de Soporte Vectorial (SVR por sus siglas en inglés) [49]. Conceptualmente funciona igual que su contraparte para clasificación, excepto que el hiperplano de separación ahora se utiliza como una superficie de regresión.

El método SVR se aplicó al pronóstico de precipitación en el Gran Chaco Argentino. Una vez obtenidas las variables predictoras para cada mes de cada grupo, la idea es utilizarlas para predecir la precipitación. Para ello se entrenaron diversos modelos de regresión por vectores de soporte.

Se probaron para cada grupo y cada mes en total 16 modelos de regresión, alternando los siguientes parámetros:

- Kernel: polinómico, sigmoide
- Grado (solo aplica a Kernel polinómico): 1; 2; 3; 4; 5; 12
- Coef: 0; 5; 10; 15; 20
- C: 1; 2; 3; 4; 5

Cabe mencionar que el Kernel lineal no se encuentra entre la lista de kernels probados ya que este es equivalente a tener un Kernel polinómico de grado 1

Al igual que con las redes neuronales, para cada modelo regresor se aplicó un criterio de selección de los mejores modelos. Para ello se calcularon las métricas de error medio absoluto (MAE) y la varianza explicada a efectos de tener resultados comparables con los derivados con los otros métodos utilizados.

3.3.3. XGBoost

Desde su presentación en el año 2014, XGBoost se convirtió rápidamente en un modelo muy popular y ha acumulado un impresionante registro de competencias ganadas. Como lo indica Didrik Nielsen en su tesis de maestría [50], en el año 2015 de las 29 competencias publicadas en la plataforma Kaggle, 17 fueron ganadas utilizando XGBoost.

La potencia de XGBoost proviene de la utilización de gradient boosting (de ahí su nombre). Esta técnica crea un modelo predictor a partir del ensamble de modelos débiles (en este caso árboles de decisión) y los generaliza a través de una función de pérdida común.

La implementación de XGBoost utilizada en esta tesis, permite definir algunos hiperparámetros que determinan el comportamiento del algoritmo. Los parámetros que fueron utilizados en este trabajo se listan a continuación:

- **Target:** Determina el valor a predecir, en este caso un número (regresión).
- **MaxDepth:** Determina la profundidad máxima de los árboles de decisión mientras más profundos son los árboles, más probabilidad de sobre-ajustar los datos.
- **Estimators:** Cantidad de árboles a utilizar.
- **RegAlpha:** Regularización L1
- **RegLambda:** Regularización L2

Este método fue aplicado al pronóstico de precipitación en el Gran Chaco Argentino. Una vez obtenidas las variables predictoras para cada mes de cada grupo, la idea es utilizarlas para predecir la precipitación. Para ello se entrenaron en total 96 modelos de XGBoost diferentes para cada par [Grupo, Mes]. Los mismos corresponden a todas las combinaciones posibles de los siguientes parámetros:

Tesis - Juan Ignacio Mazza

- Target: reg:Linear, reg:Gamma
- MaxDepth: 2; 3; 6
- Estimators: 100; 250; 500; 1000
- Reg Alpha: 0; 0,1
- Reg Lambda: 0; 0,1

Al igual que con los modelos anteriores, para cada modelo regresor se aplicó el criterio de selección de los mejores modelos calculando las métricas de error medio absoluto (MAE) y la varianza explicada a efectos de tener resultados comparables.

Se fijó un criterio de selección de las mejores redes. De cada una de las 48 redes probadas se calcularon dos métricas, el error medio absoluto (MAE por sus siglas en inglés) [51] que mide la media de los valores absolutos de los errores:

$$MAE = \frac{\sum_{i=1}^n |Y_{real(i)} - Y_{Predicho(i)}|}{n} = \frac{\sum_{i=1}^n |e_i|}{n}$$

Fórmula 2. Error absoluto medio (MAE por sus siglas en inglés)

Y la varianza explicada que indica la ganancia en términos de varianza explicada con respecto a un modelo base que predice la media.

$$Expl. Var (Y_{real}, Y_{predicho}) = 1 - \frac{Var (Y_{real} - Y_{predicho})}{Var(Y_{real})}$$

Fórmula 3. Varianza explicada

Se consideró como la mejor de las redes a la que tuviera el menor error medio absoluto y la mayor varianza explicada.

3.4. Ensamble de modelos

Sólo a modo de prueba, y teniendo en cuenta la hipótesis de que diferentes modelos pueden funcionar mejor en diferentes circunstancias, se procedió a realizar un ensamble teniendo en cuenta el mejor modelo de cada tipo (red neuronal, XGBoost y regresión de soporte vectorial).

La cantidad final de modelos a ensamblar para cada grupo y cada mes depende de las métricas individuales de los mismos, dado que para que un modelo participe del ensamble debe tener algo que aportar. Se requiere que los modelos expliquen al menos un 15% de la varianza de la precipitación para ser incluidos en el ensamble (“criterio de corte”), con el fin de que todos los modelos que lo conforman aporten al pronóstico.

Bajo estas condiciones, existen tres escenarios posibles a la hora de ensamblar los modelos:

1. **Ninguno de los modelos alcanza la varianza explicada mínima**, en cuyo caso no se generan predicciones y el sistema reporta un mensaje indicando esta situación.
2. **Sólo uno de los modelos alcanza la varianza explicada mínima**, en este caso no hay ensamble, las predicciones utilizadas son las de este único modelo.
3. **Dos o más modelos superan la varianza explicada mínima**, en este caso se realiza una suma ponderada por las respectivas varianzas explicadas. Por ejemplo, supongamos que los tres modelos tienen las siguientes varianzas explicadas: [0.276, 0.376, 0.349]. A estos valores se les aplica la función softmax, para convertirlos en valores entre 0 y 1 que suman 1, lo que da como resultado [0.314, 0.347, 0.338]. Estos son los pesos utilizados para ponderar las predicciones de cada modelo, por lo que la predicción final es de la forma:

$$\sum_{m \in \text{Modelos}} \text{Peso}_m \cdot \text{Prediccion}_m$$

Fórmula 4. Formula de ponderación de las predicciones.

La lista completa de modelos seleccionados y la eficiencia de los mismos aplicada al período de verificación (2009-2019) se detalla en el ANEXO 5. Se muestra a continuación un ejemplo para el mes de enero (M01) del grupo 1 (G1).

Se construyó un diagrama del valor de precipitación observado (valor real) vs. el valor pronosticado. Claramente la situación ideal es que todos los puntos caigan sobre la recta de identidad. En cada uno de los siguientes gráficos, los puntos verdes en el diagrama de dispersión representan el valor real y los puntos azules el valor predicho. Por otro lado, se clasifico cada pronóstico del período de verificación en las categorías antes detalladas (subnormal, normal y supranormal) definidas con los terciles de la serie observada en el período de entrenamiento.

Veamos primero las predicciones individuales. Para cada técnica se utilizó el criterio de selección para determinar el mejor modelo derivado con cada una de ellas. Dicho modelo se utilizó para pronosticar la precipitación en los años correspondientes al período de verificación.

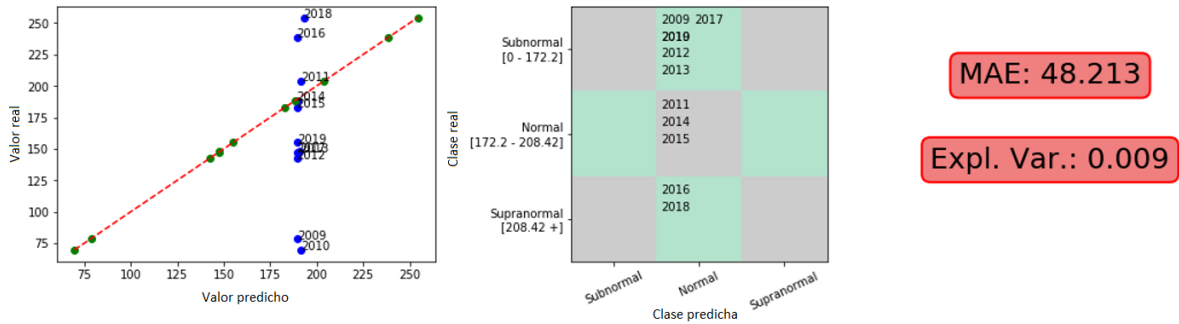


Figura 27. Predicción individual del modelo 'Red neuronal' para la precipitación de enero del G1.

Como podemos ver en la figura 27, el mejor modelo de red neuronal no logra 'aprender' a predecir los datos de este grupo en este mes y predice la media aritmética, por esto la varianza explicada es casi nula. Por lo tanto, este modelo no será incluido en el ensamble

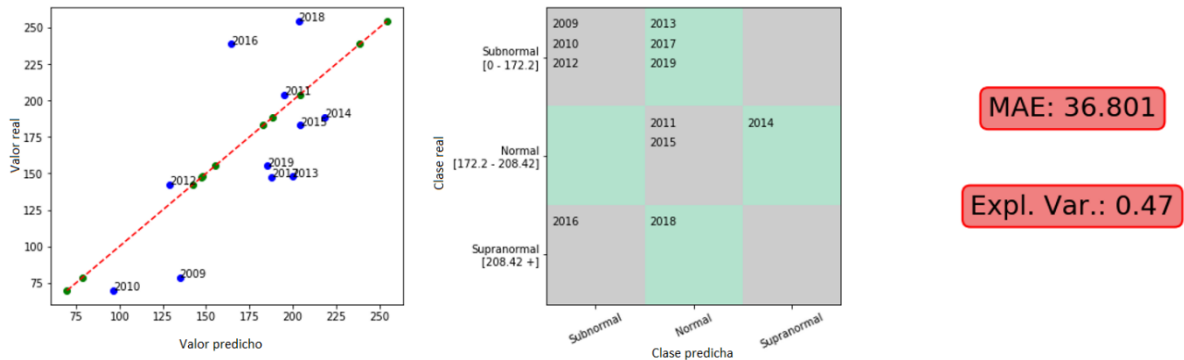
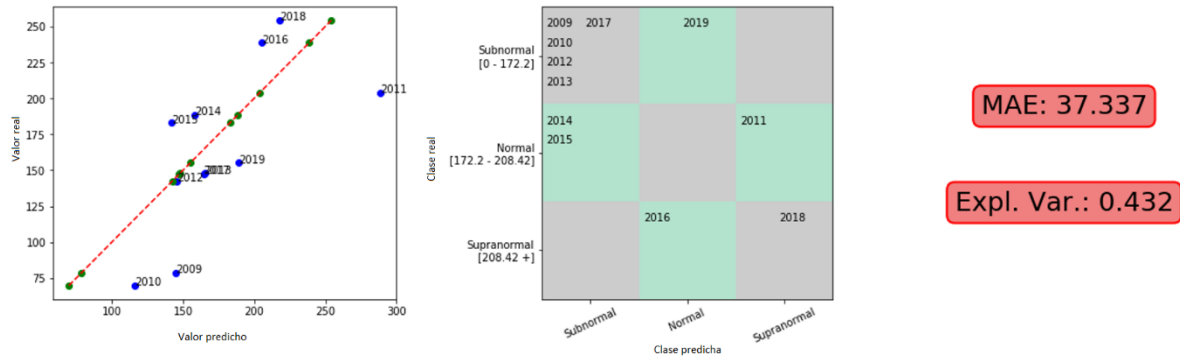


Figura 28. Predicción individual del modelo 'Regresión de soporte vectorial' para la precipitación de enero del G1.

El modelo de regresión por vectores de soporte de la figura 28 funciona mejor que la red neuronal para este grupo y mes, pero tiene problemas para aproximar los extremos (años 2016, 2018 y 2009).

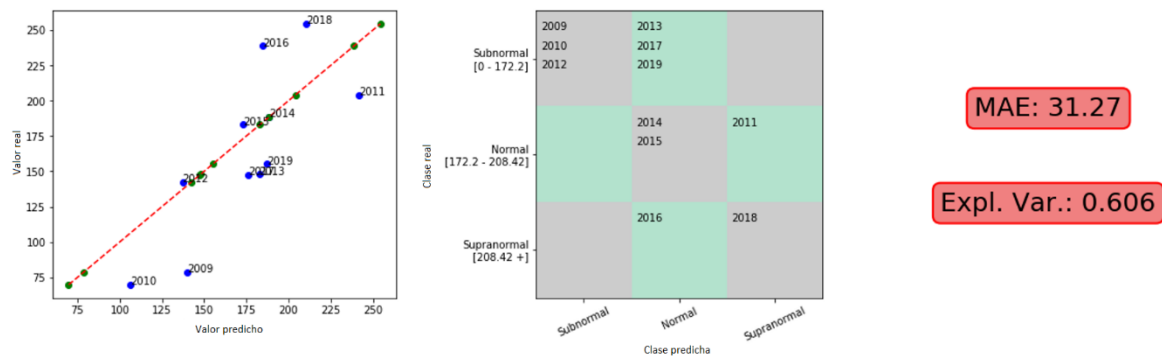


MAE: 37.337

Expl. Var.: 0.432

Figura 29. Predicción individual del modelo 'XGBoost' para la precipitación de enero del G1.

Finalmente, el modelo de XGBoost de la figura 29, logra aproximar mejor los extremos superiores (años 2016 y 2018) y la mayoría de los valores medios (años 2013, 2017 y 2019) aunque la predicción otros valores como el año 2011 empeora.



MAE: 31.27

Expl. Var.: 0.606

Figura 30. Predicción generada por el ensemble de los modelos para la precipitación de enero del G1.

El ensemble en este caso se construyó utilizando sólo los mejores modelos derivados con los métodos SVR y XGBoost, dado que la red neuronal no cumplió el criterio de selección. Como podemos ver el ensemble de los modelos XGBoost y SVR (Figura 30) logra explicar el 60% de la precipitación y clasifica correctamente 6 de los 11 años. Sin embargo, si analizamos los años incorrectamente clasificados podemos notar lo siguiente:

Los años 2013 (predicción final: 182.9 mm de lluvia), 2017 (predicción final: 176.3 mm de lluvia) y 2019 (predicción final: 187.3 mm de lluvia) quedaron próximos al límite de 172.2 mm que separa la categoría subnormal de normal para este grupo en este mes, estos errores son meteorológicamente aceptables en una región donde pueden llover hasta 250 mm en ese mes.

Tesis - Juan Ignacio Mazza

Algo similar sucede con el año 2018, cuya predicción final fue de 210.5 mm de lluvia, una diferencia de 2 mm con respecto al umbral que separa las clases normal y supranormal.

Finalmente, el año con el mayor error de predicción es el año 2011 con un valor pronosticado de 241 mm de lluvia y una diferencia con el valor observado de más de 30 mm.

Esto es consecuencia de la discretización por intervalos, se puede pronosticar una categoría diferente de la observada aun cuando haya pocos mm de diferencia entre la precipitación pronosticada y la observada.

Una vez generados los modelos de base y el ensamble, es posible obtener predicciones para cualquier grupo y mes (Anexo 5). Cabe aclarar que se debe ser cuidadoso con los pronósticos en las zonas de la región de Gran Chaco que presentan épocas sin precipitación o con precipitación muy baja. Esta situación ocurre particularmente en el noroeste argentino. La precipitación presenta una onda anual muy marcada con máximos en verano y mínimos en invierno. Esta última estación es realmente muy seca y puede presentar lluvias mensuales de menos de 5mm, por lo que se acostumbra no dar predicciones de precipitación durante estos períodos y destacarlas como zonas con “estación seca”.

3.5. La eficiencia de los modelos de pronóstico.

Si bien los modelos antes descritos son modelos de regresión, por lo que su salida es numérica, los valores de salida se discretizaron en tres categorías utilizando terciles (subnormal, normal, supranormal). Algunas medidas de eficiencia de los modelos se calculan en base a los pronósticos numéricos (varianza explicada y error cuadrático medio) y otras a los pronósticos categóricos (precision, recall, F1-Score y AUC)

Continuando con el ejemplo anterior, veamos las medidas de eficiencia del ensamble para el mes de enero (M01) del grupo 1 (G1)

El error absoluto medio, mide la media de las diferencias absolutas entre los valores predichos y los valores reales.

La varianza explicada mide la ganancia en varianza con respecto a un modelo hipotético que predice utilizando la media aritmética, a este modelo se lo conoce como modelo nulo.

Recordemos además que la varianza explicada de los modelos cumple la función de determinar si el modelo participa del ensamble.

Las métricas de Precision, recall y F1 Score requieren analizar los casos clasificados utilizando 4 variables: Verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). Estas métricas fueron originalmente concebidas para problemas binarios, sin embargo, en problemas multiclase como este, es posible calcularlas para cada clase con respecto a todas las demás (1 vs todas). El valor TP indica los casos pertenecientes a la clase en cuestión para la cual se está calculando la métrica que fueron correctamente categorizados, FP representa los casos que fueron incorrectamente clasificados como pertenecientes a la clase en cuestión cuando en realidad pertenecían a alguna de las otras dos, FN representa los casos que fueron incorrectamente categorizados como pertenecientes a alguna de las otras dos clases cuando pertenecían a la clase en cuestión, finalmente, TN representa los casos correctamente clasificados como no pertenecientes a la clase en cuestión.

Precision es la capacidad del modelo para clasificar correctamente los casos positivos, es la proporción de casos etiquetados como positivos que resulta clasificada correctamente. Su mejor valor es 1 y su peor valor es nulo. Se calcula mediante la siguiente formula:

$$Precision = \frac{TP}{TP + FP}$$

Fórmula 5. Precision

Recall es la proporción de casos positivos reales que fue correctamente etiquetado como positivo.

Su mejor valor es 1 y su peor valor es nulo. Matemáticamente, se define como:

$$Recall = \frac{TP}{TP + FN}$$

Fórmula 6. Recall

F1-score puede interpretarse como un promedio ponderado de las métricas de precisión y recall del modelo, esta métrica alcanza su mejor valor en 1 y su peor valor en 0. Matemáticamente se define como:

$$F1 - Score = \frac{TP}{TP + \frac{1}{2} (FP + FN)}$$

Fórmula 7. F1-Score

Finalmente, la métrica AUC es el área bajo la curva en un diagrama llamado diagrama ROC donde se grafican la tasa de positivos reales (casos positivos acertados sobre total de casos positivos reales, TPR por sus siglas en inglés) vs. la tasa de falsos positivos (Casos incorrectamente clasificados como positivos sobre el total de casos positivos reales, FPR por sus siglas en inglés) y mide la capacidad del modelo para pronosticar mejor que el azar (Probabilidad 0.5 de acertar la clase, definida por la diagonal). Cuanto mayor es el valor de AUC (más cercano a la unidad) lo cual indica una mayor distancia a la diagonal, mejor es el modelo. Valores menores o iguales a 0.5 indican que la predictibilidad no difiere del azar.

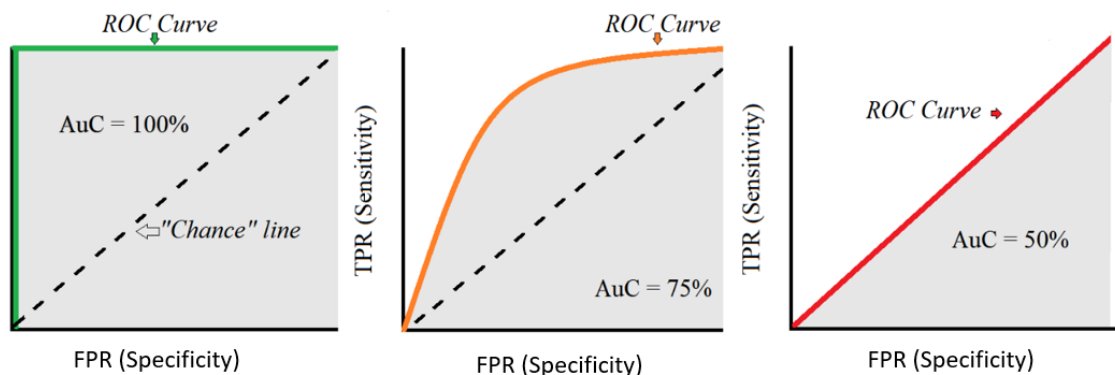


Figura 31. Ejemplo de curvas ROC y sus respectivos AUC.

La figura 31 muestra diferentes ejemplos de curvas ROC con sus respectivas áreas bajo la curva (AUC) pintadas en gris. La curva de la izquierda (en verde) es la curva optima, mientras que la curva de la izquierda, en rojo, muestra un modelo que no logra predecir mejor que el azar

Para los pronósticos de precipitación en el Gran Chaco Argentino, se consideraron los ensambles correspondientes a cada grupo y cada mes y se calcularon las medidas de eficiencia. Se detallan a continuación las medidas de eficiencia para el ejemplo del G1 y mes de enero:

MAE y Varianza explicada:

MAE: 31.27

Expl. Var.: 0.606

Figura 32. MAE y varianza explicadas del ensamble (M01 - G1)

Podemos observar que el ensamble para el mes de enero del grupo 1 explica un 60% más de varianza que el modelo nulo (predecir la media aritmética).

Precision, recall y F1-Score:

	Precision	Recall	F1-Score
Subnormal	0.67	0.5	1.0
Normal	0.44	0.67	0.33
Supranormal	0.5	0.5	0.5

Figura 33. Tabla de precisión, recall y F1-score para el ensamble (M01 - G1)

Estas métricas muestran que el modelo puede identificar correctamente el 50% de las veces las clases subnormal y supranormal (recall) y puede identificar el 67% de los casos normales. Recordemos que en el ejemplo que estamos analizando varias observaciones caen en la categoría incorrecta por muy pocos milímetros de error. Con respecto a la precisión resulta máxima en el caso subnormal (67%) mientras que es del 55% para supranormal y del 44% para la categoría normal. El F1-score es mejor para casos subnormales (100%) que para el resto (33% y 50% para categorías normal y supranormal, respectivamente). Dados estos valores, en este caso se puede afirmar que el ensamble de modelos funciona mejor para pronosticar lluvias bajas que para lluvias fuertes.

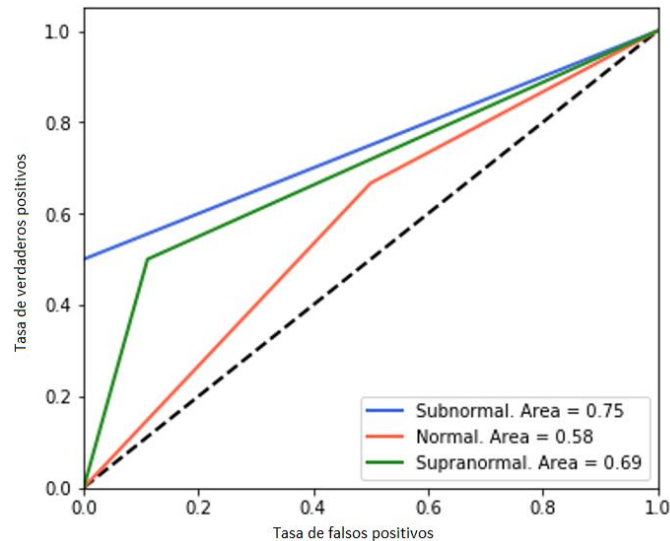
Curva ROC

Figura 34. AUC para el modelo de ensamble (M01 - G1)

De la curva ROC podemos decir que el modelo puede predecir mejor las lluvias subnormales y supranormales, cometiendo la mayor cantidad de errores en la categoría normal, esto coincide con el F1-Score del punto anterior para estas categorías.

3.6. Resultados obtenidos

3.6.1. El error de los modelos

Como pudimos observar a lo largo del presente trabajo, es posible predecir la precipitación a un mes utilizando modelos meteorológicos estadísticos. Sin embargo, deben tenerse en cuenta ciertas consideraciones:

- Los modelos no producen buenos resultados en todos los pares (Mes-Grupo), algunos de ellos no pueden ser predichos con el método sugerido en esta tesis
- Predecir precipitaciones utilizando este método para los inviernos de las regiones secas no tiene sentido desde el punto de vista meteorológico, estas zonas tienen precipitaciones mensuales de pocos mm por lo que es común no generar pronósticos para ellas y mencionar que se está produciendo la estación seca.

A continuación se detallan (Figura 35, Tabla 1) los errores absolutos medios (MAE) del ensamble para cada grupo y mes derivados de los modelos que se detallan en el ANEXO 5. Los valores sin datos son grupos y meses para los cuales no se generó una predicción debido a que ninguno de los modelos individuales cumplió el criterio de corte (explicó una varianza de la lluvia de al menos 15%). Puede observarse que algunas regiones y algunas épocas del año tienen más predictibilidad que otras.

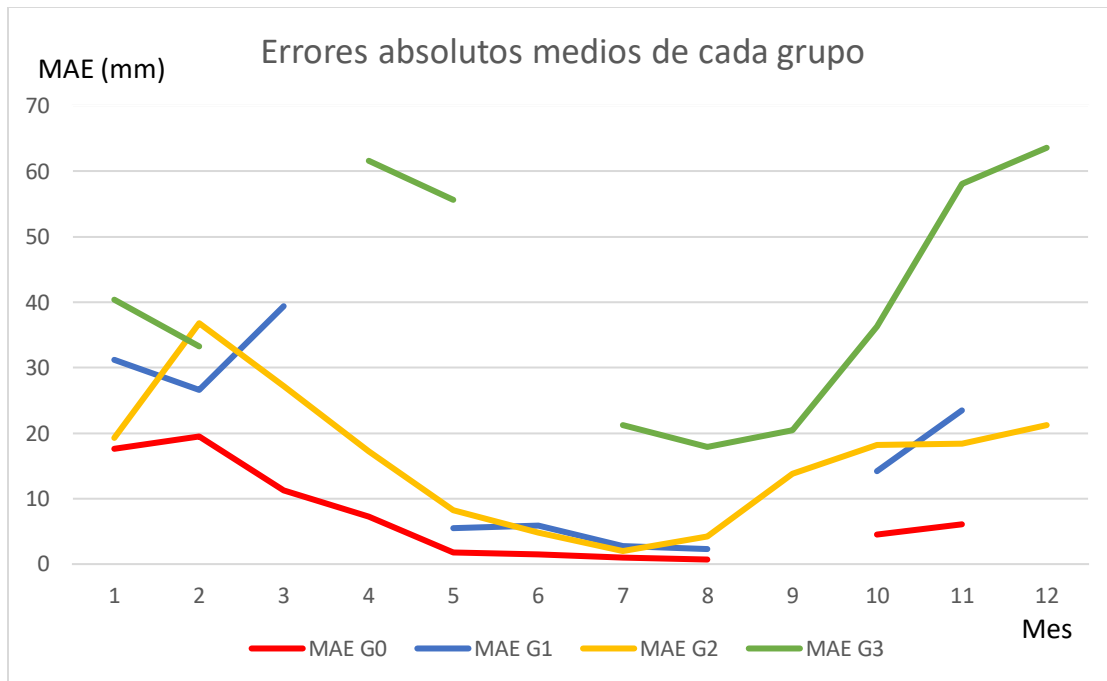


Figura 35. Errores absolutos medios para cada grupo-mes

Sin embargo, debido a que no todas las regiones tienen regímenes de precipitación comparables, resulta de interés relativizar las lluvias con respecto a los mismos, por ejemplo, un error absoluto medio de 10 mm en una zona donde llueven 20 mm por mes es un error considerable, sin embargo, el mismo error de 10 mm en una zona donde llueven 1400 mm por mes podría ser despreciable. Una manera de relativizar las precipitaciones es dividir el error absoluto medio de cada par Grupo-Mes por la mediana de las precipitaciones (Tabla 2) de dicho Grupo-Mes a lo largo de los años. La Tabla 3 y la Figura 36 muestran el error relativo derivado de este proceso, las celdas vacías pertenecen a pares Grupo-Mes para los que no se generó una predicción.

MES	MAE G0	MAE G1	MAE G2	MAE G3
1	17.6	31.2	19.3	40.4
2	19.5	26.6	36.8	33.3
3	11.3	39.4	27.2	
4	7.3		17.2	61.6
5	1.8	5.5	8.2	55.6
6	1.5	5.9	4.8	
7	1	2.8	2	21.2
8	0.7	2.3	4.2	17.9
9			13.8	20.5
10	4.5	14.2	18.2	36.3
11	6.1	23.5	18.4	58.1
12			21.2	63.6

Tabla 1. Errores absolutos medios para cada par Grupo-Mes

MES	Mediana G0	Mediana G1	Mediana G2	Mediana G3
1	58.6	188.2	116	149.4
2	43	166.9	105.6	137.6
3	27.4	129.6	94.7	139
4	11.4	48.7	53.5	135.4
5	2	12.3	15.7	79.4
6	1.5	4	5	55
7	1.8	2	3.7	29.4
8	1.4	0.8	2.4	41.2
9	1.3	4.4	21	66.8
10	6	36	58.5	137
11	16.3	86.5	87.9	161.2
12	38	142.6	110.2	127.7

Tabla 2. Medianas de precipitación para cada par Grupo-Mes

MES	MAE Relativo G0	MAE Relativo G1	MAE Relativo G2	MAE Relativo G3
1	0.30	0.17	0.17	0.27
2	0.45	0.16	0.35	0.24
3	0.41	0.30	0.29	
4	0.64		0.32	0.45
5	0.90	0.45	0.52	0.70
6	1.00	1.48	0.96	
7	0.56	1.40	0.54	0.72
8	0.50	2.88	1.75	0.43
9			0.66	0.31
10	0.75	0.39	0.31	0.26
11	0.37	0.27	0.21	0.36
12			0.19	0.50

Tabla 3. Errores absolutos medios relativizados por la precipitación mediana de cada Grupo-Mes

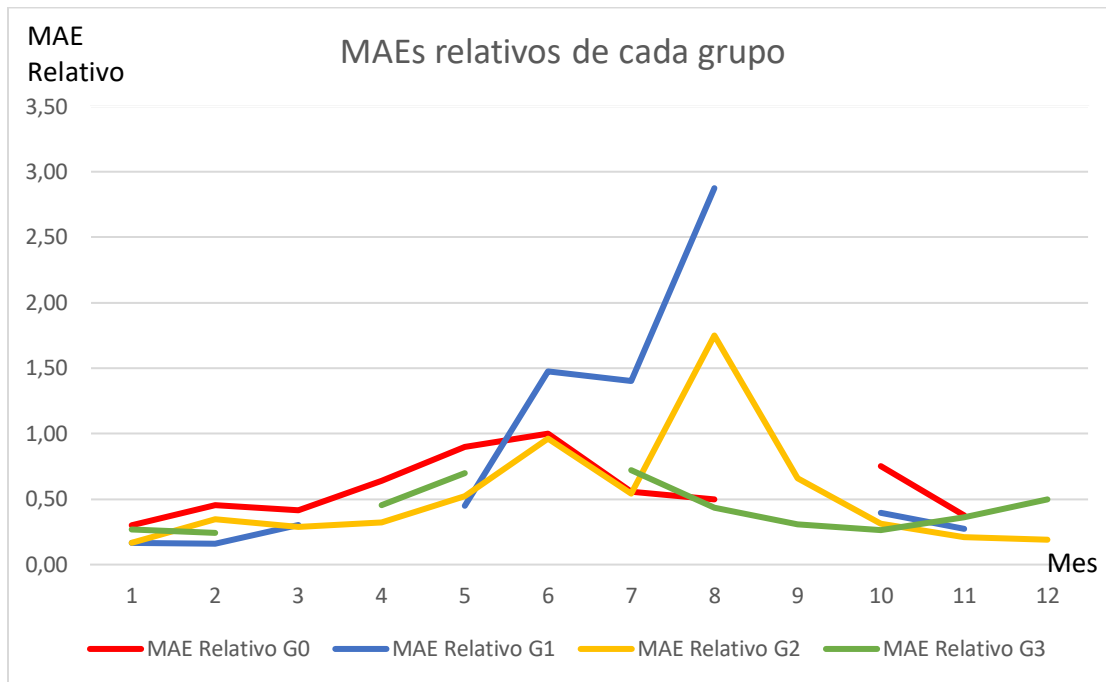


Figura 36. Errores absolutos medios relativizados por la precipitación mediana de cada Grupo-Mes

Se puede observar el aumento del error relativo en invierno en general, dado que toda la región registra menor precipitación en esa época del año. En el resto del año en que la lluvia es más alta y por lo tanto el pronóstico tiene relevancia, el error relativo se mantiene bajo. En todos los casos el error relativo obtenido es menor a una mediana, las únicas excepciones a esto son:

1. junio, julio y agosto del grupo 1, donde en estos meses secos la precipitación es tan baja (con medianas de 4 mm, 2 mm y 0,8 mm respectivamente) que resulta irrelevante incluso pronosticarla, tomemos por ejemplo agosto, este tiene una mediana de 0,8 mm, si pronosticáramos 3 mm, tendríamos un error que puede ser bajo en términos absolutos (2,2 mm) pero que resulta en un error alto en términos relativos (200% de error)
2. lo mismo sucede con agosto del grupo 2, cuya mediana es 2,4 mm. Esta precipitación es tan baja que carece de sentido pronosticar precipitaciones.

Se puede observar en la Tabla 2 que existe una relación entre los meses con precipitación baja y el alto error medio de la figura 36

3.6.2. La varianza de la precipitación explicada por los métodos

Grupo 0

El grupo 0 es una zona con una onda anual de precipitación con máximos en verano y precipitación muy baja en invierno. El modelo funciona adecuadamente en los meses más lluviosos de esta región (enero, febrero, marzo y abril), mientras que en los meses más secos (mayo, junio, julio y agosto) la predicción no es buena. No se obtuvieron modelos de predicción para septiembre y diciembre. Cabe destacar que en todos los casos el valor de MAE es inferior a 20 mm.

La tabla 4 muestra los modelos individuales (redes neuronales -NN-, regresión de soporte vectorial –SVR- y XGBoost -XGB-) que participaron del ensamble. Estos sólo fueron incluidos en el ensamble si la varianza de la precipitación que explican individualmente superaba el límite establecido de 0.15, es decir, explicaban al menos un 15% más de variabilidad que el modelo nulo que predice utilizando la media (“criterio de corte”). Nótese que algunos de los modelos resultaron en varianzas explicadas negativas, esto se debe a que estos modelos predicen peor que el modelo nulo.

Se resaltan en rojo los modelos que no superaron el criterio de corte necesario para participar del ensamble, si ninguno de los modelos de un determinado mes logro alcanzar este mínimo, no se generaron predicciones para dicho mes. Se puede observar que la técnica NN genera modelos que explican un valor relativamente alto de la varianza de la precipitación (53% en febrero, 65% en mayo y 63% en agosto y 60.8 en promedio) pero en el resto de los meses los valores de varianza explicada no alcanzan a cumplir el criterio de corte. Los métodos SVR y XGB explican en forma bastante generalizada, menor cantidad de varianza de la precipitación que NN (38.9% el SVR y 37.8% el XGB) pero existen menor cantidad de meses en que el criterio de corte no se cumple.

Mes	Grupo 0		
	NN	SVR	XGB
1	0	0.269	0.041
2	0.535	0.51	0.366
3	-0.371	0.421	-0.276
4	0.13	0.078	0.177
5	0.654	0.456	0.222
6	0.028	0.482	0.703
7	0	0.212	0.319
8	0.635	0.457	0.415
9	-0.001	0.017	-0.05
10	0.019	0.52	0.399
11	0.004	0.175	0.419
12	0	0.059	0.086
Promedio para meses que cumplen criterio de corte	0.608	0.389	0.378
Número de meses que cumplen el criterio de corte	3	9	8

Tabla 4. Varianzas explicadas de los modelos individuales del grupo 0

Grupo 1

Lo primero que destaca del grupo 1 es que 3 de los meses no generaron predicciones debido a que los modelos no lograron explicar la suficiente varianza (Tabla 5): abril, septiembre y diciembre. Sin embargo, en casi todos los meses en que se pudieron generar modelos de predicción, el MAE resultó inferior a 30 mm. Al igual que el grupo anterior, la precipitación presenta una onda anual muy marcada con mínimos en invierno por lo cual esta será considerada “estación seca”.

En el caso del grupo 1 la tabla 5 muestra las varianzas de la precipitación explicada por cada modelo, con el mismo criterio detallado en el grupo 0.

Se resaltan en rojo los modelos que no cumplieron el criterio de corte necesario para participar del ensamble, si ninguno de los modelos de un determinado mes logro alcanzar este mínimo, no se generaron predicciones para dicho mes.

En el caso del grupo 1 los distintos métodos logran explicar varianzas de la precipitación que cumplen el criterio de corte en 7 (en el caso de NN) u 8 de los 12 meses del año. Existen algunos meses como febrero, mayo, junio, agosto donde todos los métodos explican un porcentaje similar de varianza, mientras que existen otros meses donde un método funciona mejor. Por ejemplo, en julio el método SVR resulta explicar mejor la varianza de la lluvia (56,7%) que los otros dos métodos

Tesis - Juan Ignacio Mazza

(22,8 y 17,1%). Lo mismo ocurre en el mes de noviembre donde SVR explica un 58,7% de la varianza frente a 25,6 y 19,1% explicados por los otros dos métodos. La varianza promedio en los meses que cumplieron el criterio de corte es similar para todos los métodos aunque levemente inferior en el caso de NN (39.3%) en relación a SVR (49.5%) y XGB (46.6%).

Mes	Grupo 1		
	NN	SVR	XGB
1	-0.02	0.47	0.432
2	0.355	0.549	0.391
3	0.163	0.009	-0.746
4	0.007	0.07	0.122
5	0.606	0.582	0.513
6	0.397	0.226	0.33
7	0.228	0.567	0.171
8	0.746	0.457	0.818
9	0.015	0.004	-0.582
10	0.078	0.521	0.88
11	0.256	0.587	0.191
12	0	0.044	0.025
Promedio para meses que cumplen criterio de corte	0.393	0.495	0.466
Número de meses que cumplen el criterio de corte	7	8	8

Tabla 5. Varianzas explicadas de los modelos individuales del grupo 1

Grupo 2

El grupo 2 representa una zona más húmeda que los grupos anteriores, pero donde todavía domina el ciclo anual de lluvia con mínimos en invierno. Esta zona es el corazón del Gran Chaco Argentino. El MAE de este grupo es en general menor a 20 mm (excepto en febrero que supera los 30 mm), teniendo, al igual que en los casos anteriores su intervalo mínimo en los meses de menor precipitación (mayo, junio, julio y agosto).

Es importante mencionar que este es el único grupo que no cuenta con meses que no hayan podido ser pronosticados, por lo cual se puede concluir que el grupo 2 es el grupo con mayor predictibilidad.

La siguiente tabla 6 muestra la varianza de la precipitación explicada por cada método, tal como en los grupos anteriores.

Se resaltan en rojo los modelos que no cumplieron el criterio de corte necesario para participar del ensamble, si ninguno de los modelos de un determinado mes logro alcanzar este mínimo, no se generaron predicciones para dicho mes.

Claramente los métodos SVR y XGB son más eficientes en cuanto a cumplir el criterio de corte. En ambos casos, sólo 2 meses no lo cumplieron. Mientras tanto el método NN, cumple dicho criterio sólo en 5 de los 12 meses. Los meses de febrero y octubre presentan una varianza explicada de la precipitación similar usando cualquier método. Se puede observar que el método SVR explica más varianza que los otros dos, excepto en el caso del mes de mayo donde NN lo supera con un 70,5% de varianza explicada. En promedio los tres métodos explican una varianza semejante cuando se consideran sólo los meses en que se cumplió el criterio de corte. La mayor varianza promedio la explica el método SVR (48.8%). El método XGB explica una varianza promedio de 39.2%, promediando los 10 meses donde el criterio de corte se cumplió, mientras que es algo mayor en el caso de NN (42.5%) pero dicho promedio sólo pudo calcularse en base a 5 meses que cumplieron el criterio de corte.

Mes	Grupo 2		
	NN	SVR	XGB
	-		
1	0.099	0.494	0.221
2	0.454	0.419	0.423
	-		
3	0.001	0.404	-0.244
4	0.314	0.751	0.61
5	0.705	0.651	0.258
6	0.002	0.003	0.657
	-		
7	0.031	0.312	0.248
8	0.317	0.682	0.543
9	0.051	0.137	0.264
10	0.333	0.481	0.478
11	0	0.337	0.015
12	0.029	0.348	0.221
Promedio para meses que cumplen criterio de corte	0.425	0.488	0.392
Número de meses que cumplen el criterio de corte	5	10	10

Tabla 6. Varianzas explicadas de los modelos individuales del grupo 2

Grupo 3

El grupo 3 es el grupo más húmedo de todos: la precipitación es alta todo el año, aunque mayor en verano que en invierno. Para este grupo no se obtuvieron modelos de predicción para los meses de marzo y junio.

La siguiente tabla 7 muestra la varianza de la precipitación que explican los modelos individuales.

Se resaltan en rojo los modelos que no cumplieron el criterio de corte necesario para participar del ensamble, si ninguno de los modelos de un determinado mes logro alcanzar este mínimo, no se generaron predicciones para dicho mes.

Salvo en el mes de enero, el método NN explica muy poca varianza de la lluvia. Los métodos SVR y XGB parecen tener una eficiencia parecida en términos de varianza explicada. Nuevamente el método NN sólo cumple el criterio de corte en 4 de los 12 meses y además explica una varianza promedio de la precipitación que es la menos de los tres métodos (29.6%). Los métodos SVR y XGB tienen una performance similar, SVR explica un promedio de 39% de la avrianza de la precipitación y pudo ser definido en 8 meses mientras que XGB explica en promedio 36.7% de l avarianza de la precipitación en 10 meses en que cumplió el criterio de corte.

Mes	Grupo 3		
	NN	SVR	XGB
1	0.53	0.53	0.467
2	0.001	0.011	0.278
3	0.04	0.016	-0.062
4	0.058	0.232	0.295
5	0.001	0.254	0.307
6	-0.04	-0.318	0.028
7	0.247	0.002	0.483
8	0.167	0.312	0.282
9	0	0.387	0.454
10	0.009	0.487	0.28
11	0	0.516	0.479
12	0.238	0.405	0.349
Promedio para meses que cumplen criterio de corte	0.296	0.390	0.367
Número de meses que cumplen el criterio de corte	4	8	10

Tabla 7. Varianzas explicadas de los modelos individuales del grupo 3

3.6.3. Los pronósticos categóricos

Las secciones anteriores se enfocaron en analizar las performances de cada modelo para cada grupo y cada mes, veamos ahora algunas conclusiones más generales utilizando los ensambles y la métrica de área bajo la curva ROC (AUC) del ANEXO 5

Las siguientes tablas muestran las áreas bajo la curva generada por los ensambles para cada categoría pronosticada usando terciles (subnormal, normal y supranormal) para cada grupo y cada mes. Las celdas grises representan los meses para los que no se generaron modelos, mientras que las celdas con la leyenda 'N/A' representan casos donde no hubieron ejemplos dentro del tercil en cuestión, por ejemplo, no hubieron lluvias subnormales en los datos de verificación para el mes de abril del grupo 0, por lo que no se pudo computar una curva ROC y su correspondiente AUC para dicha categoría, esto se debe a que en las regiones más secas del país los intervalos de cada categoría son demasiado pequeños (por ej. 0-5 mm de precipitación) y podría suceder que no haya datos del conjunto de test que caigan en todos los intervalos.

Mes	AUCs Grupo 0			AUCs Grupo 1		
	Subnormal	Normal	Supranormal	Subnormal	Normal	Supranormal
1	0.86	0.73	0.5	0.75	0.58	0.69
2	0.5	0.65	0.94	0.62	0.72	0.8
3	0.83	0.28	0.54	0.5	0.64	0.69
4	N/A	0.65	0.65			
5	N/A	1	1	0.5	0.75	0.92
6	N/A	1	1	N/A	0.62	0.62
7	N/A	0.75	0.75	N/A	0.83	0.83
8	N/A	1	1	N/A	0.75	0.75
9						
10	N/A	0.68	0.68	0.72	0.54	0.75
11	0.73	0.53	0.67	0.67	0.72	0.83
12						
Promedio	0.73	0.727	0.773	0.627	0.683	0.764

Mes	AUCs Grupo 2			AUCs Grupo 3		
	Subnormal	Normal	Supranormal	Subnormal	Normal	Supranormal
1	0.75	0.79	1	0.5	0.66	0.78
2	0.67	0.22	0.55	0.88	0.63	0.5
3	0.67	0.66	0.9			
4	0.5	0.46	0.88	0.8	0.83	0.68
5	0.5	0.75	0.68	0.6	0.54	0.73
6	N/A	0.54	0.54			
7	N/A	0.75	0.75	0.43	0.65	0.73
8	N/A	0.89	0.89	0.55	0.58	0.68
9	0.71	0.73	0.75	0.8	0.64	0.5
10	0.62	0.58	0.73	0.88	0.62	0.75
11	0.62	0.44	0.83	0.62	0.7	0.75
12	0.82	0.61	0.69	0.5	0.75	0.73
promedio	0.651	0.618	0.766	0.656	0.660	0.683

Tabla 8. Curvas AUC de los ensambles para todas las categorías de todos los grupos

Como podemos ver, de las 111 curvas ROC computadas (excluyendo los grupos y meses que no generaron modelos predictores y las categorías sin ejemplos representadas por 'N/A') solo 27 (24%) resultaron por debajo de un umbral establecido arbitrariamente en 0.65 (valores sombreados en rojo en la tabla 8). Sólo unos pocos casos tuvieron AUC menor a 0.5 lo que indica que la predictibilidad no supera al azar.

En todos los grupos y categorías el promedio de AUC supera 0,6 lo que indica que los modelos generados producen pronósticos mejores que el azar. El grupo 0 tienen valores superiores a los otros grupos, pero con muchos meses donde los modelos no pudieron definirse. En general, no se observan diferencias marcadas en la efectividad de los otros grupos. Sin embargo, algo importante de mencionar, es que algunas categorías tienen más predictibilidad que otras. Es evidente al observar la tabla 8 que la categoría supranormal tiene en general valores más altos de AUC que las otras.

Las tablas siguientes (tablas 9, 10, 11 y 12) muestran los estadísticos Recall, Precision y F1-score individuales y el promedio de estos para todos los meses, para el modelo ensamble de cada grupo. Las filas grises representan meses para los que en ese grupo no se generó un modelo, las celdas sin valor que contienen el carácter "-", indican que para el grupo y mes en cuestión, no había observaciones en el test de validación pertenecientes a la etiqueta (subnormal, normal, supranormal)

GRUPO 0

Mes	Subnormal			Normal			Supranormal		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
1	0,8	1	0,67	0,73	0,67	0,8	0	0	0
2	0	0	0	0,67	0,8	0,57	0,86	1	0,75
3	0,8	0,67	1	0	0	0	0,33	0,33	0,33
4	-	-	-	0,71	0,62	0,83	0,5	0,67	0,4
5	-	-	-	1	1	1	1	1	1
6	-	-	-	1	1	1	1	1	1
7	-	-	-	0,95	1	0,9	0,67	0,5	1
8	-	-	-	1	1	1	1	1	1
9									
10	-	-	-	0,8	0,86	0,75	0,57	0,5	0,67
11	0,73	0,67	0,8	0,29	0,5	0,2	0,5	0,33	1
12									
PROM	0,5825	0,585	0,6175	0,715	0,745	0,705	0,643	0,633	0,715

Tabla 9. Estadísticos Precision, Recall y F1-Score para el modelo ensamble del grupo 0

Resulta evidente al observar los promedios de la tabla 9 para cada etiqueta (subnormal, normal y supranormal) que la clase subnormal tiene menor predictibilidad que las otras en este grupo, a su vez, ambas clases, normal y supranormal muestran buenos valores de precisión, recall y F1-Score (valor 1 en muchos casos), lo que indica que el modelo de ensamble del grupo 0 es capaz de reconocer y pronosticar estas categorías de precipitación.

Finalmente, cabe destacar que la baja performance de la categoría subnormal es relativa, dado que de los cuatro meses que tuvieron observaciones en esta categoría, tres de ellos obtuvieron buenos resultados, sin embargo, el promedio final se vio afectado por los resultados de febrero donde los valores de todos los estadísticos fue 0.

GRUPO 1

Mes	Subnormal			Normal			Supranormal		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
1	0,67	0,5	1	0,44	0,67	0,33	0,5	0,5	0,5
2	0,4	0,25	1	0,44	1	0,29	0,75	0,6	1
3	0	0	0	0,62	1	0,44	0,5	0,5	0,5
4									
5	0	0	0	0,29	1	0,17	0,91	0,83	1
6	-	-	-	0,71	0,83	0,62	0,5	0,4	0,67
7	-	-	-	0,83	1	0,71	0,8	0,67	1
8	-	-	-	0,95	1	0,9	0,67	0,5	1
9									
10	0,77	0,83	0,71	0,33	0,33	0,33	0,67	0,5	1
11	0,5	0,33	1	0,44	1	0,29	0,8	0,67	1
12									
PROM	0,39	0,318333	0,618333	0,561111	0,87	0,453333	0,677778	0,574444	0,852222

Tabla 10. Estadísticos Precision, Recall y F1-Score para el modelo ensamble del grupo 1

Al observar los resultados promedio de la tabla 10, sucede algo similar al grupo anterior, la categoría subnormal vuelve a mostrar la peor performance en términos de precisión, recall y F1-Score. Por otro lado, la categoría Normal muestra un Recall promedio de 0,87, lo que significa que en líneas generales el modelo de ensamble de este grupo acierta esta categoría el 87% de las veces, sin embargo, la baja precisión de esta categoría nos indica que el modelo predice normal mas de lo que debería, en otras palabras, solo el 56,11% de las veces que el modelo predice normal acierta, esto genera que el F1-Score final sea relativamente bajo (0,45).

El caso inverso sucede con la clase supranormal, el modelo de ensamble tiene mayor precisión que recall, el 67,77% de las veces que el modelo predice supranormal acierta, pero solo identifico correctamente el 57% de los casos reales de esta categoría, aun así, el F1-Score de esta categoría casi duplica al de la categoría normal.

GRUPO 2

Mes	Subnormal			Normal			Supranormal		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
1	0,67	0,5	1	0,73	1	0,57	1	1	1
2	0,5	0,33	1	0	0	0	0,55	0,5	0,6
3	0,5	0,33	1	0,6	0,75	0,5	0,5	1	0,33
4	0	0	0	0,4	0,5	0,33	0,75	1	0,6
5	0	0	0	0,77	1	0,62	0,57	0,5	0,6
6	-	-	-	0,75	0,75	0,75	0,33	0,33	0,33
7	-	-	-	0,95	1	0,9	0,67	0,5	1
8	-	-	-	0,88	0,78	1	0,67	1	0,5
9	0,57	0,67	0,5	0,67	0,75	0,6	0,67	0,5	1
10	0,4	0,25	1	0,33	0,5	0,15	0,73	0,8	0,67
11	0,71	0,83	0,62	0	0	0	0,8	0,67	1
12	0,8	0,8	0,8	0,5	0,5	0,5	0,5	0,5	0,5
PROM	0,461111	0,412222	0,657778	0,548333	0,6275	0,493333	0,645	0,691667	0,6775

Tabla 11. Estadísticos Precision, Recall y F1-Score para el modelo ensamble del grupo 2

Al observar los promedios de la tabla 11, podemos observar que la clase subnormal posee la menor predictibilidad en términos de precisión y recall, nuevamente el modelo muestra problemas para predecir las precipitaciones de esta categoría, sin embargo esto es esperable debido a las pocas observaciones del test de validación que existían para la categoría subnormal de este grupo (ver ANEXO 5), por ejemplo, en el mes de febrero existieron solo 3 observaciones subnormales (2009, 2018 y 2019), y el modelo predijo correctamente solo 1 de ellas, esto significa que acertó 1/3 de los casos obteniendo un recall de 33%, resulta evidente entonces, que en los meses donde hay pocos ejemplos los errores se vuelven mas costosos, en este caso particular, con solo 3 observaciones, cada año mal clasificado implica una perdida de recall de 33%.

Finalmente, la categoría supranormal es la que muestra mayor predictibilidad en este grupo, el modelo tiene una precisión de 64% (64% de las veces que asigna esta categoría acierta) y logro identificar un 69% de los casos pertenecientes a esta categoría.

GRUPO 3

Mes	Subnormal			Normal			Supranormal		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
1	0	0	0	0,67	0,57	0,8	0,5	1	0,33
2	0,75	1	0,6	0,67	0,67	0,67	0	0	0
3									
4	0,75	0,6	1	0,57	1	0,4	0,57	0,5	0,67
5	0,4	0,33	0,5	0,33	0,33	0,33	0,73	0,8	0,67
6									
7	0	0	0	0,5	0,67	0,4	0,67	0,75	0,6
8	0,33	0,25	0,5	0,44	0,67	0,33	0,57	0,5	0,67
9	0,86	1	0,75	0,4	0,5	0,33	0	0	0
10	0,86	0,75	1	0,4	0,25	1	0,6	1	0,43
11	0,4	0,25	1	0,25	1	0,14	0,67	0,5	1
12	0	0	0	0,29	1	0,17	0,73	0,67	0,8
PROM	0,435	0,418	0,535	0,452	0,666	0,457	0,504	0,572	0,517

Tabla 12. Estadísticos Precision, Recall y F1-Score para el modelo ensamble del grupo 3

La tabla 12, perteneciente al grupo 3 nos muestra que este grupo es el que mas problemas de predictibilidad tiene, con valores medios globales inferiores a los de los otros grupos y en varios casos por debajo del 50%. Ninguna categoría perteneciente a este grupo muestra una contundente mejor performance que las otras.

CAPÍTULO IV

CONCLUSIONES

4.1. Conclusiones

En este trabajo de tesis se ha modelado el pronóstico de precipitación mensual en la región del Gran Chaco Argentino. Este pronóstico es relevante pues representa una herramienta más para la toma de decisiones en una zona donde las actividades productivas agrícolas son la principal fuente de recursos y donde las inundaciones por desborde de ríos es muy frecuente, generando un perjuicio notable para la población.

Las principales conclusiones derivadas de este trabajo son:

- El área de estudio se regionalizó en 4 grupos con diferente régimen de precipitación, utilizando el método de mapas auto-organizativos o redes neuronales de Kohonen (SOM). Se observó que la lluvia presenta un ciclo anual con mínimos en invierno que se profundiza hacia el oeste, mientras que en la región este la precipitación es alta todo el año y el ciclo anual es menos marcado.
- La aplicación de metodologías de pronóstico de precipitación requirió definir predictores en el mes previo al de pronóstico. Para ello se utilizaron datos de variables meteorológicas y oceánicas globales que se correlacionaron con la precipitación producida al mes posterior. Las zonas con correlación significativa con 95% de confianza fueron utilizadas para definir predictores con la condición de que los mismos pudieran explicarse físicamente.
- Los modelos de pronóstico se diseñaron utilizando técnicas de data mining: NN, SVR y XGBoost. Se probaron 48 arquitecturas de redes neuronales, 16 modelos de regresión de soporte vectorial y 96 modelos de XGBoost. Para cada grupo y para cada mes, se consideró el mejor modelo obtenido con cada método observando que tuviera el menor error absoluto medio (MAE) y que explicara la mayor cantidad de varianza de la precipitación.
- Se consideró un ensamble de los mejores modelos con la condición de cumplir el criterio de corte (explicar más de un 15% de la varianza de la precipitación). De esta forma resultaron meses donde el ensamble no pudo definirse o bien donde estuvo compuesto por sólo 1 o 2 métodos.
- Todos los modelos fueron entrenados para el período 1980-2008 y verificados en 2009-2019.
- Los modelos no producen buenos resultados en regiones y meses donde la precipitación es muy baja. Tal es el caso de los meses invernales en el noroeste. De hecho los centros de pronóstico rotulan esta situación como “estación seca” y no emiten pronóstico.

Tesis - Juan Ignacio Mazza

- El error medio relativo (MAE relativizado con la mediana de precipitación) es mayor en invierno que en verano en toda la región dado que el ciclo anual muestra mínimos en invierno.
- En el caso del Grupo 0 las NN logran explicar el 60.8% de la varianza promedio de la lluvia pero sólo pudieron diseñar modelos en 3 meses. SVR y XGBoost explicaron menor varianza pero se obtuvieron modelos para 8 y 9 meses.
- En el caso del Grupo 1 SVR y XGBoost explicaron varianzas similares y mayores al 46% y pudieron generar modelos en 8 meses, superando a la eficiencia de NN.
- En el caso del Grupo 2 la mejor técnica fue SVR explicando un 48.8% de la varianza en promedio y definiendo modelos para 10 meses. También resultó la mejor técnica en el caso del Grupo 3, explicando un 39% de la varianza en promedio y definiendo modelos en 8 meses. En este caso la eficiencia de XGBoost fue similar.
- NN tienen consistentemente la peor performance en todos los grupos. Sin embargo, existen casos (aislados) donde las mismas obtienen resultados excelentes llegando a explicar más del 70% de la variabilidad con respecto al modelo nulo (agosto del grupo 1 o mayo del grupo 2). Se observa que en términos generales, SVR explica mayor proporción de la varianza que los otros dos métodos.
- Se realizaron pronósticos por categorías: subnormal, normal y supranormal. Los diagramas ROC de los ensambles para cada grupo y promedio sobre los meses indica que en todos los casos los modelos generados mejoran un pronóstico al azar. El estadístico AUC superó 0,6 en todos los casos.
- La categoría supranormal fue mejor pronosticada que las otras en todos los grupos.

La conclusión general es que estas técnicas de data mining han sido útiles para procesar grandes cantidades de datos y generar modelos relativamente eficientes que pueden ser agregados a otros modelos (estadísticos o determinísticos) para generar ensambles como los que se utilizan normalmente para el pronóstico de precipitación.

4.2. Ideas para futuros desarrollos

Como ideas para futuros desarrollos, se proponen:

- La exploración de diferentes arquitecturas de los modelos en esta tesis utilizados
- La exploración de nuevos modelos
- La exploración de nuevas variables predictoras
- La aplicación de esta técnica a otras regiones de la Argentina
- La incorporación de otras fuentes de datos que pudieran ser relevantes

4.3. Repositorio de archivos

Todo el material utilizado en esta tesis de maestría, papers, scripts, shape files, datasets, etc. Se encuentra disponible en el siguiente repositorio de github:

<https://github.com/JuanMazza85/TesisUBA>

4.4. Referencias

- [1] A. Mary, "Deterministic chaos , fractals , and quantumlike mechanics in atmospheric flows," *Can. J. Phys.*, vol. 68, no. 9, 1990.
- [2] V. Gonzalez, M. and Barros, "Aspectos estadísticos del ciclo anual de precipitación y sus anomalías en Argentina subtropical," *Meteorologica*, 21, n° 1 y 2. *Meteorologica*, pp. 15–26, 1996.
- [3] R. B. y M. C. Vicente Barros, Moira Doyle, Marcela González, Ines Camilloni, "Revision of the south american monsoon system and climate in subtropical south america south of 20°S." *Meteorologica*, 27, pp. 35–58, 2002.
- [4] Barnston. A.. Kumar. A.. Goddard. L. and Hoerling. M, "Improving seasonal prediction practices through attribution of climate variability," *BAMS*, pp. 59–72, 2005.
- [5] K. A, "On the interpretation and utility of skill information for seasonal climate predictions," *Mon. Wea. Rev*, vol. 135, pp. 1974–1984, 2006.
- [6] Leetmaa. A, "Seasonal Forecasting. Innovation in practice and institutions," *BAMS*, vol. 84,

Tesis - Juan Ignacio Mazza
pp. 1686–1691, 2003.

- [7] G. Coelho, C. Stephenson, D. Balmaseda, M. Doblado Reyes, F. And Oldenborge, “Towards an integrated seasonal forecasting system for South America,” *J. Clim.*, vol. 19, pp. 3704–3721, 2005.
- [8] S. Goddard, L. Barnston, A. and Mason, “Evaluation of the IRI’s ‘net assessment’ seasonal climate forecasts 1997-2001,” *BAMS*, pp. 1761–1781, 2003.
- [9] Barreiro. M, “Influence of ENSO and the South Atlantic Ocean on climate predictability over Southeastern South America.,” *Clim. Dyn.*, vol. 35, pp. 1493–1508, 2010.
- [10] A. Nobre, C., Marengo, J., Cavalcanti, I., Obregon, G., Barros, V., Camilloni, I., Campos, N. and Ferreira, “Seasonal to decadal predictability and prediction of South America Climate,” *J. Clim.*, pp. 5988–6004, 2005.
- [11] C. J. C. Reason, “Subtropical Indian Ocean SST dipole events and southern South African rainfall,” *Geophys. Res. Lett.*, vol. 28, no. 11, pp. 2225–2227, 2001.
- [12] D. Karoly and V. Dayton, *Meteorology in Southern Hemisphere*. 1999.
- [13] J. E. Oliver and J. J. Hidore, *Climatology, an introduction*. 1984.
- [14] J. P. Peixoto, *Physics of climate*. 1984.
- [15] J. R. Holton and R. Dmowska, *El Niño, La Niña, and the southern oscillation*. Academic press, 1989.
- [16] M. L. Salby, *Physics of the Atmosphere and Climate*. 2012.
- [17] K. E. Trenberth and T. J. Hoar, “The 1990-1995 El Niño-Southern Oscillation event: Longest on record,” *Geophys. Res. Lett.*, vol. 23, no. 1, pp. 57–60, 1996.
- [18] C. F. Ropelewski and M. S. Halpert, “Global and Regional Scale Precipitation Patterns Associated with the El Niño-Southern Oscillation,” *Mon. Weather Rev.*, vol. 115, no. 8, pp. 1606–1626, 1987.
- [19] C. Vera, G. Silvestri, V. Barros, and A. Carril, “Differences in El Niño Response over the Southern Hemisphere,” *J. Clim.*, vol. 17, no. 9, pp. 1741–1753, 2004.
- [20] A. M. Grimm, “Interannual climate variability in South America : impacts on seasonal precipitation , extreme events , and possible effects of climate change,” *Stoch. Environ. Res.*

- [21] S. E. Marcela H. González, Eugenia M. Garbarini, Alfredo L. Rolla, “Meteorological Drought Indices: Rainfall Prediction in Argentina,” *Handb. Drought Water Scarcity*, vol. 29, pp. 540–567, 2017.
- [22] G. M. y R. A. Garbarini E, Skansi M, “ENSO influence over precipitation,” *Argentina, Adv. Environ. Res.*, vol. 52, no. 7, pp. 223–246, 2016.
- [23] “ENSO cycle.” [Online]. Available: https://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensocycle/enso_cycle.shtml.
- [24] N. H. Saji and P. N. Vinayachandran, “A dipole mode in the tropical Indian Ocean,” *Nature*, vol. 401, no. September, pp. 360–363, 1999.
- [25] S. Chan, S. Behera, and T. Yamagata, “Indian Ocean Dipole influence on South American rainfall : Climatic impacts of Indian Ocean dipoles, El Nino-Southern oscillation, and their interaction with the monsoon systems in the Asia-Oceania region,” *Geophys. Res. Lett.*, vol. 35, no. 14, 2008.
- [26] T. Taschetto, A. S. y Ambrizzi, “Can Indian Ocean SST anomalies influence South American rainfall,” *Clim Dyn*, vol. 38, pp. 1615–1628, 2012.
- [27] T. Behera, S. K., y Yamagata, “Subtropical SST dipole events in the southern Indian Ocean,” *Geophys. Res. Lett.*, vol. 28, no. 2, pp. 327–330, 2001.
- [28] T. Gissila, E. Black, D. I. F. Grimes, and J. M. Slingo, “Seasonal forecasting of the ethiopian summer rains,” *Int. J. Climatol.*, vol. 24, no. 11, pp. 1345–1358, 2004.
- [29] “Dipolo del indico.” .
- [30] V. E. Kousky, “Precipitation and atmospheric circulation anomaly patterns in the South American sector,” *Rev. Bras. Meteorol.*, vol. 3, pp. 199–206, 1988.
- [31] C. F. Gan, M. A., Kousky, V. E., & Ropelewski, “The South America Monsoon Circulation and Its Relationship to Rainfall over West-Central Brazil,” *J. Clim.*, vol. 17, no. 1, pp. 47–66, 2004.
- [32] M. A. Gan, V. B. Rao, and M. C. L. Moscati, “South American monsoon indices,” vol. 223, no. July 1979, pp. 219–223, 2006.

- [33] M. Gonzalez, J. Nery, and V. Barros, "The relation between tropical convection in South America and the end of the dry period in subtropical Argentina," *Int. J. Climatol.*, vol. 18, no. 15, pp. 1671–1687, 1998.
- [34] V. Barros and M. Gonzalez, "Climate variability over subtropical South America and the Southamerican monsoon: a review," vol. 27, pp. 33–57, 2002.
- [35] J. M. Thompson, D. W. y Wallace, "Annular modes in the extratropical circulation. Part I: month-to-month variability," *J. Clim.*, vol. 13, no. 5, pp. 1000–1016, 2000.
- [36] J. A. Marengo *et al.*, "Recent developments on the South American monsoon system," *Int. J. Climatol.*, vol. 32, no. 1, pp. 1–21, 2012.
- [37] G. E. Silvestri and C. S. Vera, "Antarctic Oscillation signal on precipitation anomalies over southeastern South America," *Geophys. Res. Lett.*, vol. 30, no. 21, pp. 1–4, 2003.
- [38] M. Gonzalez, "Some indicatros of interannual rainfall variability in Patagonia (Argentina)," *Clim. Var. - Reg. Temat. patterns*, vol. 6, pp. 133–161.
- [39] E. M. Garbarini, "Algunos indicadores para la predicción estadística de la precipitación estacional en Argentina. Tesis de Licenciatura," Departamento de Ciencias de la Atmósfera y los Océanos, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires., 2016.
- [40] A. Garbarini, EM, González, MH and Rolla, "The influence of Atlantic High on seasonal rainfall in Argentina," *Int. J. Climatol.*, vol. 39, no. 12, pp. 4688–4702, 2019.
- [41] P. Oliveri, "La influencia de los océanos cercanos sobre la precipitación y temperatura estacionales en Argentina. Tesis de Licenciatura en Ciencias de la Atmósfera, Unversidad de Buenos Aires.," 2018.
- [42] R. AL Gonzalez, MH, *Agricultural Research Updates*, vol. 27. 2019.
- [43] Imme Ebert-Uphoff_ and Kyle Hilburn, "Evaluation, Tuning and Interpretation of Neural Networks for Working with Images in Meteorological Applications," *Bull. Am. Meteorol. Soc.*, 2020.
- [44] D. Rostkier-Edelstein *et al.*, "High-resolution forecasts of seasonal precipitation: a combined statistical-dynamical downscaling approach," *Ann. EGU*, 2010.
- [45] G. T. Diro, A. M. Tompkins, and X. Bi, "Dynamical downscaling of ECMWF Ensemble seasonal forecasts over East Africa with RegCM3," *J. Geophys*, 2012.

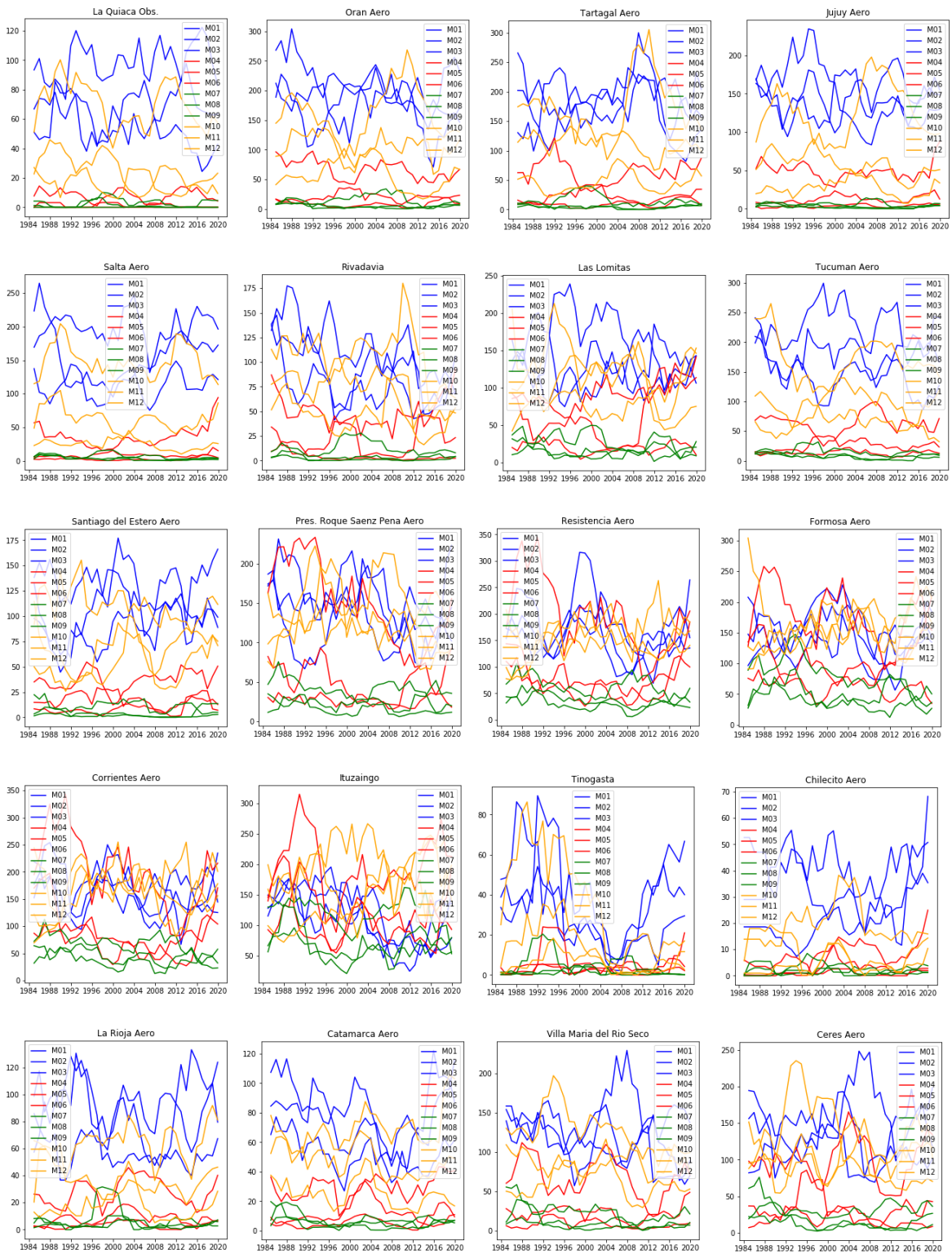
- [46] L. Yuan, L. Guihua, W. Zhiyong, and H. H., "High-Resolution Dynamical Downscaling of Seasonal Precipitation Forecasts for the Hanjiang Basin in China Using the Weather Research and Forecasting Model," *J. Appl. Meteorol. Climatol.*, vol. 56, pp. 1515–1535, 2017.
- [47] Kalnay et al, "The NCEP/NCAR 40-year reanalysis project," 1996. [Online]. Available: <https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html>.
- [48] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.
- [49] H. Drucker *et al.*, "Support Vector Regression Machines," *NIPS'96 Proc. 9th Int. Conf. Neural Inf. Process. Syst.*, vol. 1, pp. 155–161, 1996.
- [50] D. Nielsen, "Tree Boosting With XGBoost," 2016.
- [51] C. Res, C. J. Willmott, and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Clim. Res.*, vol. 30, no. 1, pp. 79–82, 2005.

ANEXOS

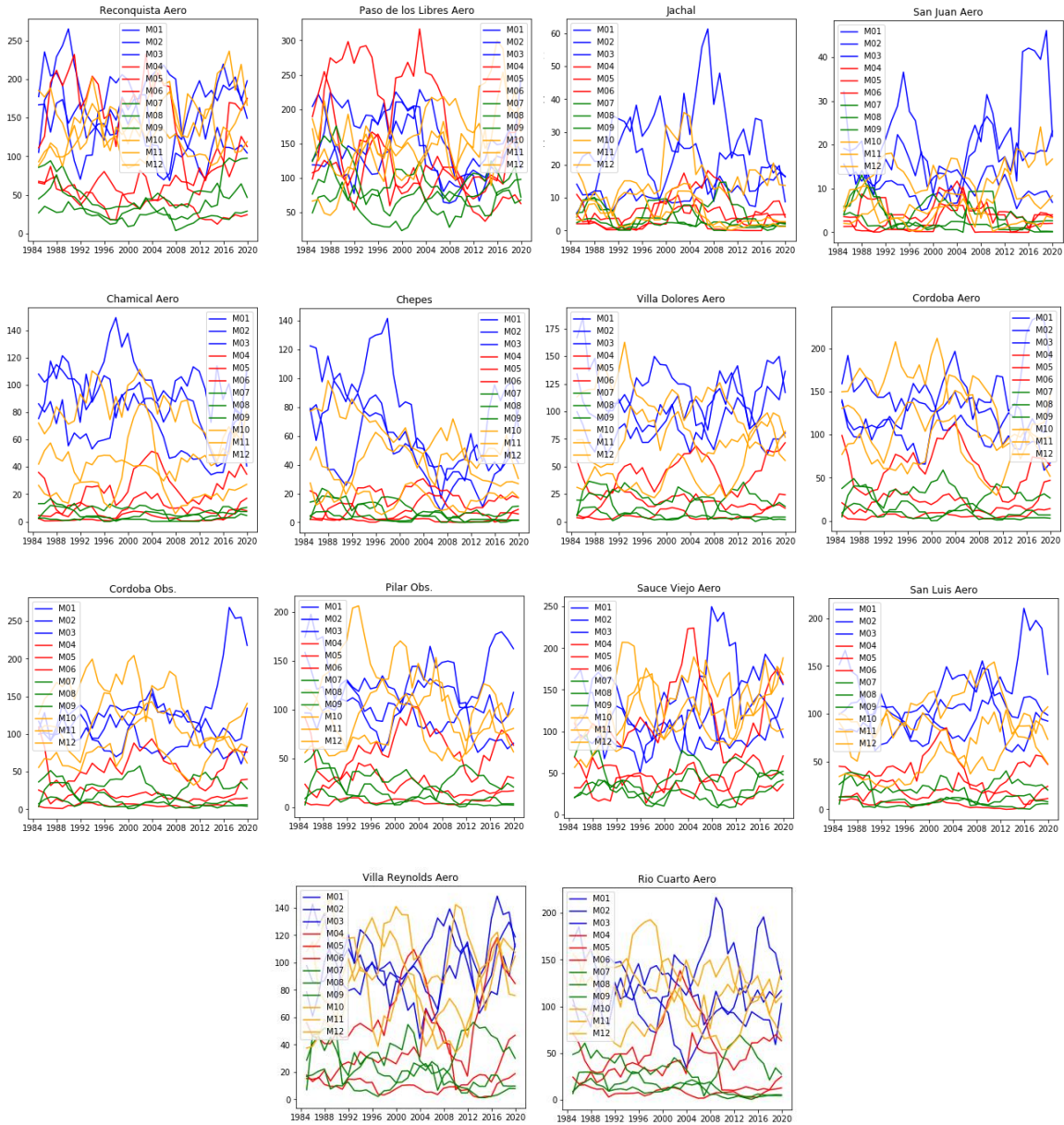
ANEXO 1 - Listado de estaciones y grupos

ID	Nombre	Grupo	Longitud	Latitud	Elevación
87007	La Quiaca Obs.	0	-65,6	-22,1	3459
87016	Oran Aero	1	-64,32	-23,15	357
87022	Tartagal Aero	1	-63,82	-22,65	450
87046	Jujuy Aero	1	-65,08	-24,38	905
87047	Salta Aero	1	-65,48	-24,85	1221
87065	Rivadavia	2	-62,9	-24,17	205
87078	Las Lomitas	2	-60,58	-24,7	130
87121	Tucuman Aero	1	-65,1	-26,85	450
87129	Santiago del Estero Aero	2	-64,3	-27,77	199
87148	Pres. Roque Saenz Pena Aero	3	-60,4	-26,75	93
87155	Resistencia Aero	3	-59,05	-27,45	52
87162	Formosa Aero	3	-58,23	-26,2	60
87166	Corrientes Aero	3	-58,77	-27,45	62
87173	Ituzaingo	3	-56,67	-27,58	72
87211	Tinogasta	0	-67,57	-28,07	1201
87213	Chilecito Aero	0	-67,43	-29,23	945
87217	La Rioja Aero	0	-66,82	-29,38	429
87222	Catamarca Aero	0	-65,77	-28,6	454
87244	Villa Maria del Rio Seco	2	-63,68	-29,9	341
87257	Ceres Aero	2	-61,95	-29,88	88
87270	Reconquista Aero	3	-59,7	-29,18	53
87289	Paso de los Libres Aero	3	-57,15	-29,68	70
87305	Jachal	0	-68,75	-30,23	1175
87311	San Juan Aero	0	-68,42	-31,57	598
87320	Chamical Aero	0	-66,28	-30,37	461
87322	Chepes	0	-66,58	-31,33	658
87328	Villa Dolores Aero	2	-65,13	-31,95	569
87344	Cordoba Aero	2	-64,2	-31,3	474
87345	Cordoba Obs.	2	-64,18	-31,4	425
87349	Pilar Obs.	2	-63,88	-31,67	338
87371	Sauce Viejo Aero	2	-60,82	-31,7	18
87436	San Luis Aero	2	-66,35	-33,27	713
87448	Villa Reynolds Aero	2	-65,38	-33,73	486
87453	Rio Cuarto Aero	2	-64,23	-33,12	421

ANEXO 2 – Series temporales de precipitación mensual acumulada para cada estación (en mm)



Tesis - Juan Ignacio Mazza



ANEXO 3 – Series medianas de precipitación acumulada (en mm) para cada grupo de estaciones

Grupo 0

YEAR	M01	M02	M03	M04	M05	M06	M07	M08	M09	M10	M11	M12
1979	128.3	54.7	18	35.6	0	0.8	1.3	0	1.3	6.8	23.4	87.5
1980	38.2	67	46.4	12.6	0.1	3.2	0	0	0	22	20.3	31
1981	78.6	40.5	35.4	32.6	0.1	0	1.7	6.8	2.4	14	15.5	39.4
1982	38.2	33	31.3	18.5	0	1.6	0.1	0	12.1	0	26.8	36.4
1983	58.6	30.3	3.7	4.6	2.6	4.2	4.4	3.8	5.6	12	18.8	41.6
1984	108	33.4	77	6	0.1	2	0	0	13.4	0.7	8	39.7
1985	90.5	117.3	23.6	0.7	0	0	7.6	14	0.2	0.6	72	54.1
1986	23.1	62.8	29.5	1.5	0	0	11.1	3.8	14.4	16	7	55.6
1987	108.6	23	19	5.3	15.9	0	0	0	1.3	13.6	15	35
1988	81.5	73.9	30	2.2	3.5	0	0	0	0.6	0	5.4	16.4
1989	54	59.5	34.5	2.1	3.5	0	0	0	1.5	3	22.4	43.5
1990	44.2	58.3	26.5	13.8	2	0	7.3	0	2	44.3	43.8	28.8
1991	78	80.3	63.1	11.4	0.5	3.5	0	0.5	0	16.4	18.2	50
1992	77.5	57.8	20.1	62	1	0	0.5	9.5	3	0	46.1	79.7
1993	107.6	42	45.7	0	0	0	1.2	0	4.7	2.5	8.6	23
1994	33.5	70	8.2	6.2	0	0	2.5	0	0	11.3	46.4	55.1
1995	35.7	32.6	31.2	0	0.7	0	0	0	12.1	0	39	14.8
1996	87.9	26	37	4	3	4	0	0	16.6	2.7	16	57
1997	100.5	30.5	25	0	8	0	0	1	2	3	12.5	42
1998	32.5	56.2	27.4	15.5	11	0	0	7.2	5	20.3	13	17
1999	61	56	61	13	4	0.9	5.8	0	3	40.5	25	68.8
2000	82	41.5	44.4	25.7	15	9.9	10	0	0	23	27.3	27.4
2001	59.1	22	49	17.6	11.9	11.4	0.1	0	15	14	17	62.8
2002	37.3	62.2	27.2	20	3.6	0	5	0	0	11.7	2	69
2003	8	31.9	38.4	16	2	0.2	0	0	0	1	1	5
2004	52	30.3	25	1.2	16	0	0	0	1.1	4	9	19
2005	31.9	70.2	20.4	19.1	0.3	0	0	4	5	2	13.3	38
2006	47.2	40.6	5.6	20.8	0.1	0	0	0	0	2	16.3	40.3
2007	49.7	5.1	26.3	6	8.8	0.4	1	2	1	4	27	28
2008	74.1	66	55	2.6	2.5	0	0	0	7	10	23.8	63.5
2009	61.9	5.4	39	0	4.3	0	0	0	0	0	5.2	28
2010	39	40	2	3	12	0	5	0	1	4	10	27
2011	24.1	88	14	12	0	0	3	0	0	12.1	20	30.7
2012	29.1	61	5	18.7	0	0	0	0	5.6	1	7	2
2013	47	15	9	0.5	0	0.3	0	1.2	1	14	5.7	51.8
2014	22	157	23	10.8	3.1	0.3	2	0	0.2	19	11.9	35.4
2015	71	57.3	59	0.3	0	0	0	6	0.8	6	33.7	40.1
2016	62	50	22	28	5	22.5	0	0	0.1	3	15	30
2017	49	43	28	14	4.3	0.1	0	0.4	5.1	4.3	3	43.5
2018	82	36	27	12.5	1	0	6	0	3	30	27.5	21
2019	73.5	23	48	20	5.2	0	0	0	0.3	8	25	16.3

Grupo 1

YEAR	M01	M02	M03	M04	M05	M06	M07	M08	M09	M10	M11	M12
------	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Tesis - Juan Ignacio Mazza

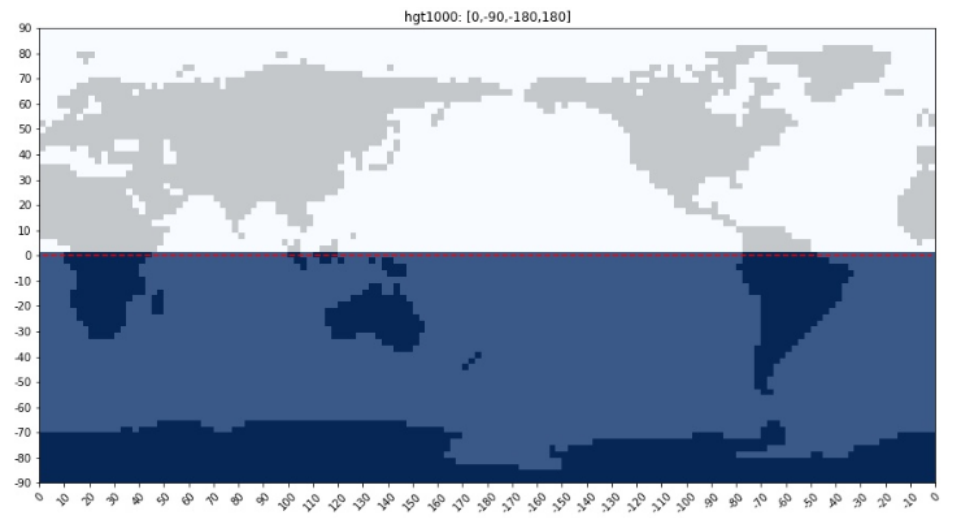
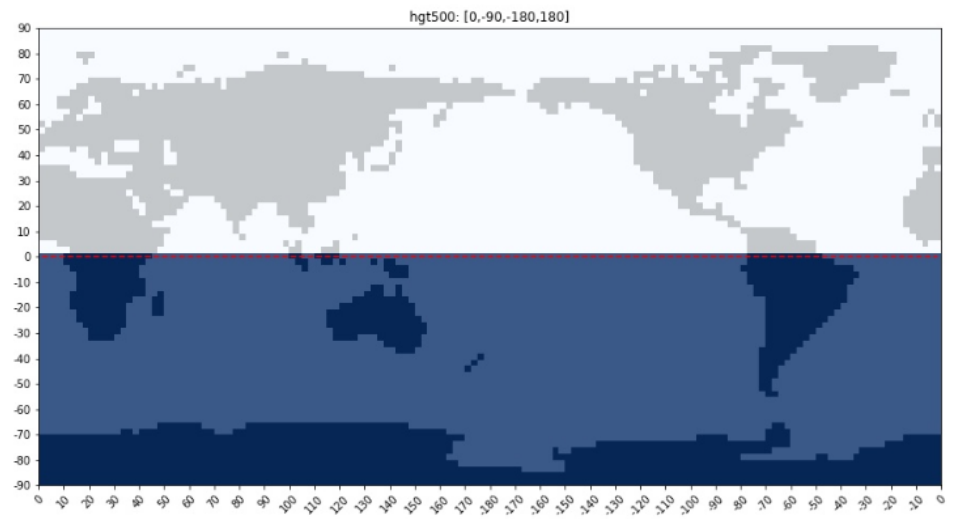
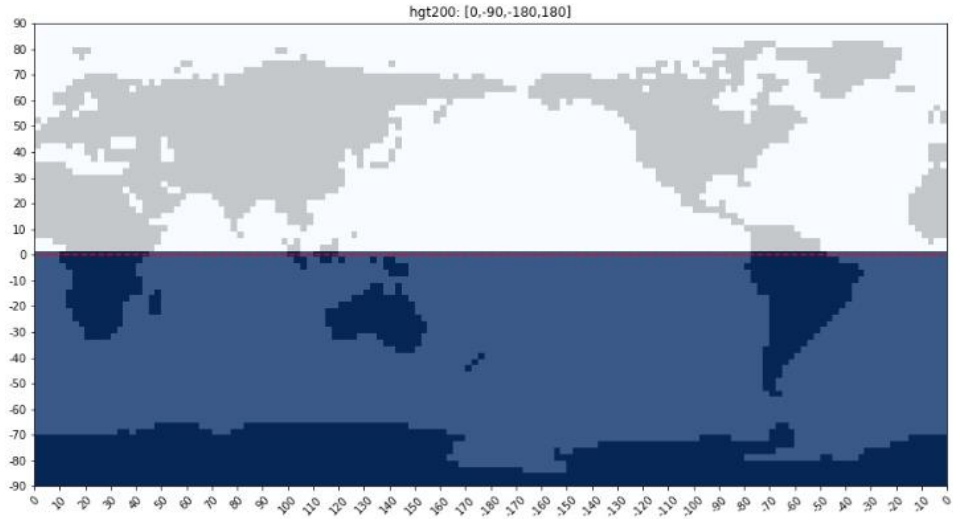
1979	217.8	115.1	163.3	62.3	1.2	1.6	3.6	6.3	7.3	39.3	132.4	157.2
1980	145.1	140.2	210.3	57.1	12.2	22.4	0	9.8	0	43.6	74.8	95.7
1981	328.7	268.5	108.1	122.8	1.2	0.7	2.2	8.9	1.6	22.1	104.8	109.2
1982	172.1	168.6	189.1	53.3	0.9	9.8	6.8	5.2	8.9	18.4	55.1	144.8
1983	190.6	198	66.6	57.6	19.9	7.1	13.3	4.8	4.1	47.5	66.8	128.7
1984	174.5	203.8	299.9	23.4	9.7	16.7	4.3	13.7	2.2	35.3	66.9	230.8
1985	214.9	243	125.3	81.4	6.3	3.8	15.8	16.4	11.8	61.3	122.1	142.6
1986	146.9	141.8	86.2	40.6	12.4	6.1	3.9	24.3	10.9	55.2	132.1	185.5
1987	287.2	60.5	79.6	72	12.3	0.2	1.3	1.3	0.5	11.6	105.6	206.3
1988	241.5	131.8	167.9	54.5	11.9	0.1	13.7	0.6	0.4	4.1	33.2	134.7
1989	87.8	54.7	253.5	58.3	8	24.9	4.6	0	5.7	23.1	72.5	136.4
1990	167.7	165.5	145.3	97.2	18.6	2.1	0.4	2.6	2.2	46.9	86.5	164.8
1991	221.5	224.9	221.6	75	19.5	5.3	2	4.2	9	47.1	163.4	123.8
1992	185.9	205.9	103.7	48.3	9.5	3.6	7.5	20.2	24.5	25.2	123.3	161.6
1993	218.7	80.6	162.9	33.1	3.7	0.3	3.6	1.1	2.1	41.9	85.7	117.3
1994	194.6	210.1	40.7	26.6	33.9	0.8	2	1	14	43.8	93.1	134.3
1995	217.8	137.7	121.8	5.6	22.4	3.1	1.3	0.2	18.5	11	56.5	33
1996	188.2	125.8	91.2	33.4	30.3	5.6	0	0	32	61.3	100.7	188.5
1997	201.5	248.2	107.8	33.6	8.3	0.7	1.7	2	6.5	14	47.3	54.1
1998	172.6	61.9	118.8	43.7	5.1	4.2	3	13.8	1.2	95.5	102.6	135.4
1999	234.9	119	267.6	35.4	32.5	8.2	8	0	13.2	59.3	33.6	75.2
2000	186	166.9	187.7	64.5	15.7	9	8	0	0.2	61.2	159	112.2
2001	257.4	168	136.8	81.2	7.8	12.3	0	0.9	27.8	42.7	97.9	137
2002	156.3	184.7	100.5	24.6	26.8	7.2	1.5	0	1	101	109.8	211.4
2003	203.2	124	146	31.1	16.2	7.9	0.6	0.4	3.7	72.6	66.6	190.7
2004	111.8	171.9	127.9	100.5	20	4	0.2	0.2	23.2	20.2	126	148.4
2005	83.3	139.9	140.8	61.7	3.8	1.6	1	0	12.3	8.8	114.7	218.5
2006	185	141.4	114.4	108.6	1.3	5.3	0	0	0	32.6	63.8	266
2007	306.4	178.8	157.4	32.8	13.5	1.7	0.5	0	7	36.1	54.6	189.6
2008	253	213.8	192.7	48.1	2.9	0.3	0	0.8	1	29.3	93.4	208.5
2009	78.6	122.6	153.4	48.7	24.4	0	0	0	0	0.1	110.6	212.2
2010	69.7	251.2	133.3	43.5	26.1	2.8	11	0.8	4	8.5	23.2	135.6
2011	204.1	254.7	90	46.4	6.5	8.6	13	0.1	5	12	37.2	162.4
2012	142.4	172	99.6	107.9	4.6	0.5	0.8	0	0.6	21.5	94.8	111.2
2013	147.8	112.5	61.1	8.8	6.6	21	1	0	1.8	42.3	42	161.4
2014	188.4	206.2	118.1	25.2	35.4	14.4	1.5	0	14.3	50.8	55.2	143.7
2015	183	195.2	129.6	85	13	3	9	21.6	4.4	12.7	62.4	93.4
2016	238.8	227.3	84.2	42.6	19.7	21.1	5	10	4.2	36	132.8	92
2017	147.4	137	209.5	38.6	7	3.9	5	0.8	5.2	25.4	39.7	141
2018	254.4	134.5	58.4	63.2	30.6	0.9	10.9	0	12.3	125.1	70	156.8
2019	155.2	138	147	101.1	33	10.3	0.8	0	0.3	16	131.1	91.5

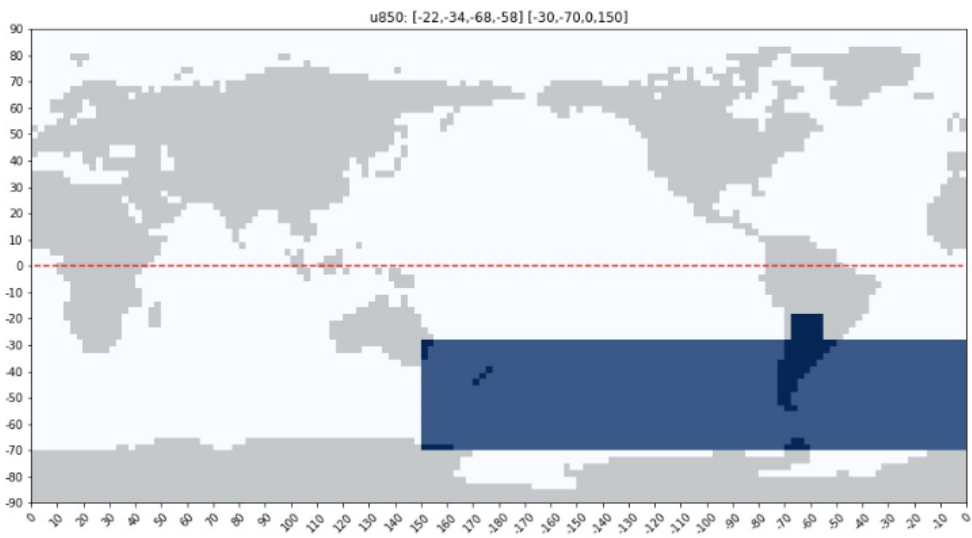
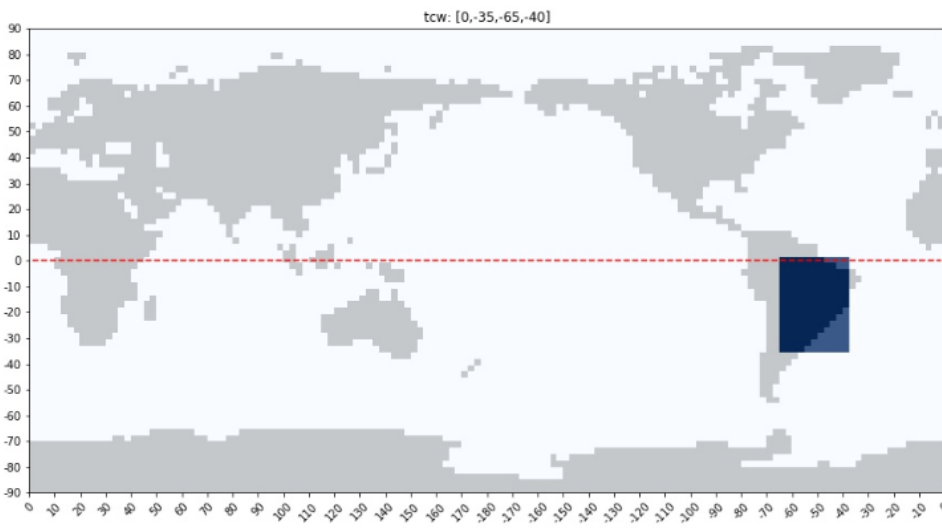
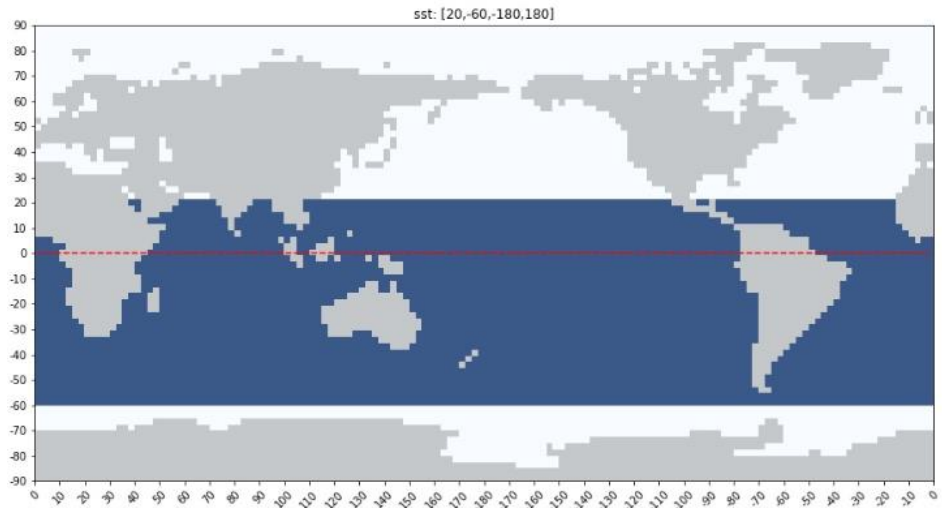
YEAR	M01	M02	M03	M04	M05	M06	M07	M08	M09	M10	M11	M12
1979	138.6	162	37.3	79.6	3.9	48.1	9.1	15.1	25.8	31.4	97.2	131.6
1980	47.6	140.9	141.2	107.3	33	13.6	0	1.3	12.1	62.7	112.4	132.9
1981	211.6	157	90.4	91.7	38.4	0.4	1.7	2	3.3	48.6	138.4	67.9
1982	83.9	105.6	113.6	73.2	8.5	11.8	6.6	2.6	71.6	27.1	79.8	81.8
1983	195.1	115.5	61.2	45.1	33.7	3.7	11.7	28.3	21.3	82.2	72.5	98.7
1984	141.7	117.2	173.9	24.2	9.6	10.4	3.1	7.2	80.8	34.9	135.5	180.5
1985	142.6	98.6	41.5	43	5	3.2	111.4	27	32.5	89.1	84.5	110.3
1986	113	83.3	61.5	34.2	9.5	8.1	23.5	20.8	37.6	60.6	82.8	100.2
1987	134.7	89.7	131.7	53.5	35.6	0.8	6.2	2.9	4.7	54.8	80.2	139.3
1988	124.2	94.7	129.6	17.1	3.8	0.5	5.1	0.3	25.2	18	57.9	75.1
1989	67.8	100.5	115.5	53.2	40.2	20.8	0	2.4	18.9	34	79.2	135.3
1990	138.3	113	116.8	38.1	14.2	1	19.2	0	9.5	85.2	74	132.5
1991	105.1	105.4	121.6	67.6	37.6	18.8	6.1	28.8	12.1	65.9	75	196.2
1992	86.5	128.7	132.6	76.4	17.8	3.7	6.5	22.6	52.2	65.1	116.8	219.6
1993	100.6	60.5	60.3	55.2	26.8	0.3	1.9	2	13.5	34	153.4	98.5
1994	131.5	87.2	94.7	46.2	22.9	5.5	3.3	26.7	6.3	75.2	92.5	71.7
1995	116.8	106.5	57.6	14.5	23.9	0.9	0.8	0	7.4	47.9	126	52
1996	161.2	115.5	82.4	112.7	14.5	5.7	0	0	39.5	60.6	78	161.5
1997	123	98	69	11.8	2.5	10.6	11.5	1.2	42.6	71	97	228.7
1998	92.3	149.8	75.6	63	32	18.2	3.2	7	27.2	45	107	64.4
1999	153.8	58.5	159.7	110	9.2	13.7	5.8	3	21	120.8	91.5	147.5
2000	138.9	107.6	115.1	123.5	47.2	5	10.3	1.7	5.5	103.4	140	96.8
2001	136	57	135.7	111.8	14.8	8.2	1.2	17.2	78	93.9	87.5	110.2
2002	115.5	83.3	132.1	75.2	17	2	27	16.5	5.1	84	104.7	125.3
2003	99.9	50.4	140.8	79.6	10.4	5	16	2.9	3	40.2	45.6	119
2004	59	61.8	115	50.2	33	0	21	0.1	4	57.3	86.8	145.6
2005	133.2	86.1	133.6	40	6.8	5.3	11	9	11.2	56	104	93
2006	148	83.3	84	93.1	2.2	6	0.4	0.3	5	58.5	102	175.9
2007	103.6	124	204.1	45.8	9.5	3.1	2.6	0.2	40.5	54	33.7	91.8
2008	162.7	122.1	132	13.1	3.9	1	0	2	39	44.4	87.9	79
2009	91	84.3	72.2	7	15	0	9.3	0	22.4	10.8	48	157.8
2010	70	116.9	86.1	47.7	33	2.3	0	0.5	62	46.1	71	58.5
2011	116	133	76	35.9	5.1	17.3	3.7	1	12	88.4	113.9	42.3
2012	103	110.3	58	65.1	14.7	0	0.2	22.5	55	118.8	76.7	83
2013	68	94.5	74.4	35.5	18.4	1.6	0	0.6	5	102	75	90.5
2014	84.5	161.3	97.7	98	24	2.9	2.3	0	37.8	55.3	101	108.4
2015	120	288	66.3	51.6	15.7	1.8	6	26.3	11.2	68	113.3	102.4
2016	111.8	192	71.3	105.7	8.2	35	2.8	1.3	1	90.6	71.7	133.7
2017	79	123.4	91	73.7	33.3	10	8	4.4	39.5	45	52.1	111.2
2018	81.9	52	11.7	44.5	75	4	2.4	2.8	23	80.9	131	79.1
2019	185.1	43	131	80.8	28.2	7.2	1.1	0	6.1	34.5	91.3	114

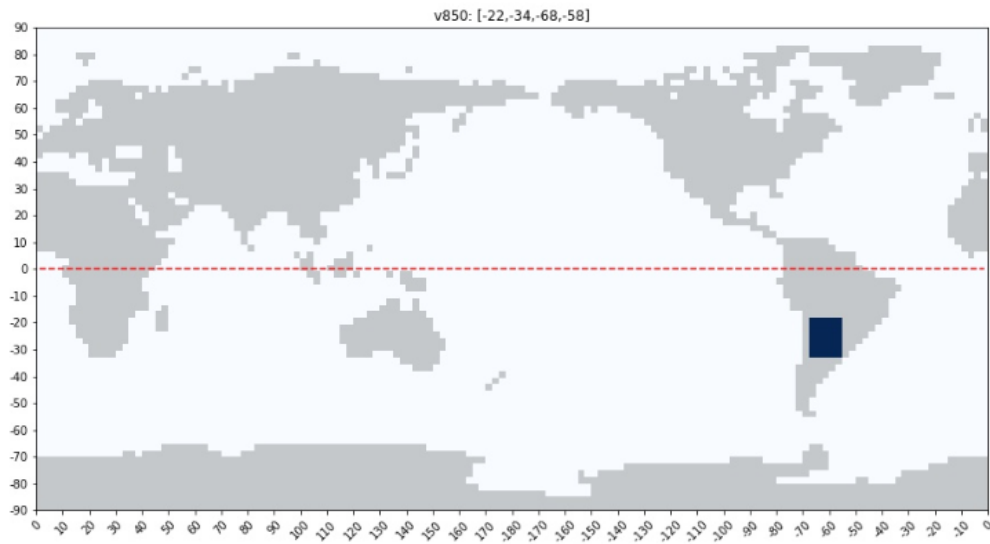
Grupo 3

YEAR	M01	M02	M03	M04	M05	M06	M07	M08	M09	M10	M11	M12
1979	49.6	123.5	131.8	124.2	79.4	55.2	29.4	63.5	107.1	156.1	112.7	247.7
1980	92.1	78.7	101.5	114.5	109.4	51.3	8.4	57.5	64	139.7	183.9	92.5
1981	275.1	233.1	123.6	130.2	110	63.6	44.7	45.2	53.8	46.9	171.4	98
1982	96.6	176.7	72.3	80.2	75.9	175.7	28.5	57.5	166.1	61.7	363.4	129.6
1983	224.8	137.6	153	208.9	203.6	25.4	75.1	12.2	23.5	115.5	109.7	51.6
1984	248.1	176.6	283.2	120.2	176.2	88.4	22.7	17.5	105.9	134.9	182.3	74.3
1985	72.1	216.5	171.5	335.4	184.2	30.2	65.1	76.8	115.1	74.7	19.8	82.8
1986	72.5	149.5	356.6	441	168.3	165.8	45.5	29.3	151.5	144.3	221.4	107.3
1987	183	185.5	82.4	231.5	68.2	74.4	119.6	41.2	30	70.3	172.9	112.9
1988	210.3	74.8	93.9	148.1	8	29.2	8.9	33.1	66.8	87.1	90.9	90.5
1989	97.5	74	226.8	263.9	10.3	75.5	45.4	98.9	106.7	131.5	99	127.7
1990	129.5	88.9	135.9	362.2	60.8	70.6	41.6	53	58.9	275	161.2	184.8
1991	121.7	88.7	35.5	150	217	86.1	29.8	9.5	76.5	93	132.7	208.2
1992	107.5	192.9	211.7	216.5	81.1	100.5	42	57.9	69.1	231	127.6	177.8
1993	190.8	17.4	177.4	135.3	44.8	44.6	36	19.7	60.4	152.3	270.5	137.8
1994	149.9	229.5	161.2	129.9	107.9	78	34	53	56.9	157.2	200	73.5
1995	160.6	202.5	185.5	73.7	80.3	6.5	31.8	6.8	48.5	60.6	71.2	65.7
1996	202.4	225.3	185.2	328	29	10.6	2	35.4	112.2	198.3	148.4	240
1997	78.5	226.1	60	67.3	87.5	27.7	26.5	21	73.4	223.3	217.6	204.4
1998	276	248	238	401.9	40.5	61.5	49.6	43.8	60.1	85	144.1	121.5
1999	99	164.6	139	116.4	94	44.1	35.5	3	22.4	64.3	59.9	96
2000	185.5	168.5	83.5	135.4	93.3	48.2	5.4	42.5	34.5	213.2	166.5	99
2001	195.2	112.8	158.3	154.1	10.3	142	21.1	64	69.6	150	163	61.5
2002	224.2	107.9	272	313.7	79.9	41	103.4	52.1	100	154.4	181.6	262.5
2003	119	107	134.2	118.1	9	19.1	5.9	63.1	29.2	136	174.4	205.5
2004	42.6	97	76	138	15	85.3	25.7	22	89.4	195.5	192.5	168
2005	69.3	42.5	125	163	119.8	69.3	3.4	33.7	48.5	81.4	112	95.5
2006	88	31.7	165.5	101.2	32	72.9	10	20	57.8	222.6	110	205
2007	204.9	155.2	201.5	132.5	25.6	16.5	5.9	8.5	116.5	216.2	87.3	164
2008	82.9	107	43.7	65.2	18	71	19.2	33.1	82	193.5	57	45
2009	118	137	19	32	58.8	21.6	26.5	25.3	38	52.2	430.1	210.5
2010	178	147.4	135	54.5	172.9	7.6	110	43	100	45.8	47.1	181.3
2011	134.1	178.7	54	121.6	66.8	45.2	47	2.2	69	144.5	206	55
2012	51	62.3	65.1	157.4	44.6	25.6	13.5	93.8	23.4	180	89.7	282.7
2013	149.4	104	169.2	183	78.4	74.5	39.7	16	36.4	112.6	235.3	59
2014	177.3	170.8	227.3	117.5	66	57	61.4	10.5	88.7	61.3	139.2	265.4
2015	162.7	143.5	69	70.9	130	61.7	6.1	73.9	12	156.4	225	308.9
2016	179	119.2	130	231.8	13.2	40	25.5	69	14.6	223	104	146.5
2017	95.3	141	177.4	311.2	219.7	55	5.2	64.7	72.1	90.4	72.4	54.6
2018	261.1	28	182.5	39.3	183.7	30	15.5	20.5	99	113	246.1	213.4
2019	352.2	125	167.3	217.5	285.3	14.7	71	45.1	16	137	196	56

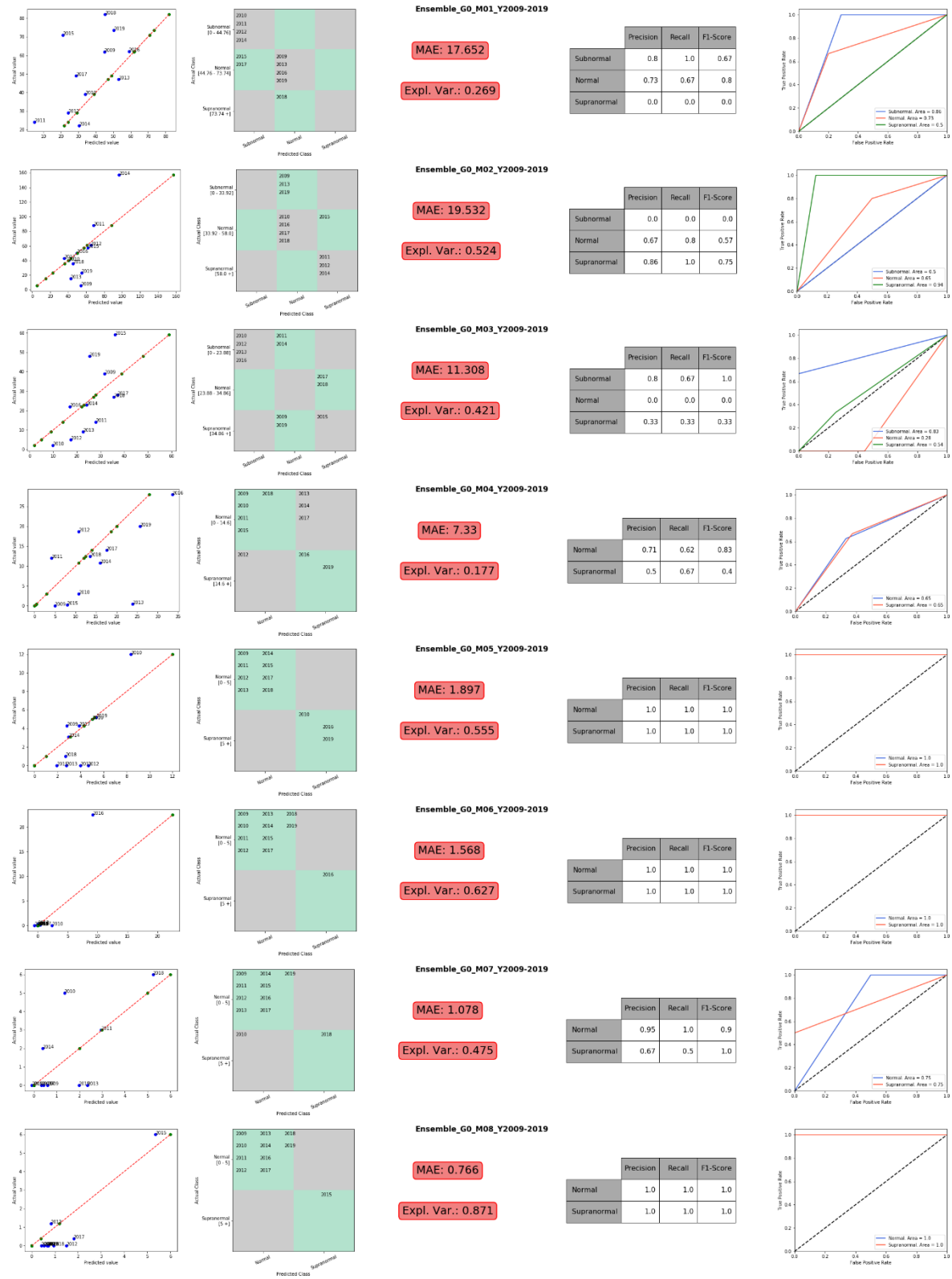
ANEXO 4 – Máscaras de filtrado

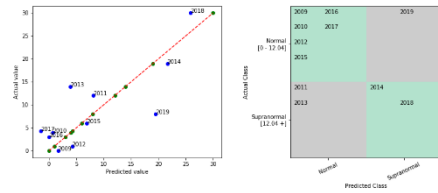






ANEXO 5 – Resultados de los modelos



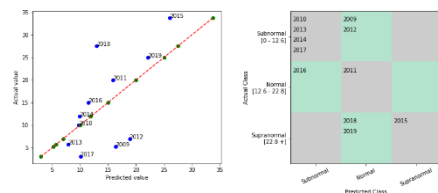
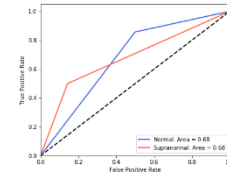


Ensemble_G0_M10_Y2009-2019

MAE: 4.567

Expl. Var.: 0.592

	Precision	Recall	F1-Score
Normal	0.8	0.86	0.75
Supranormal	0.57	0.5	0.67

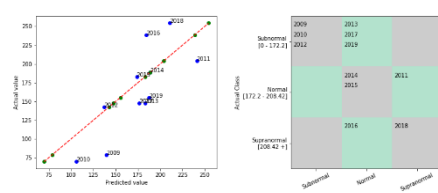
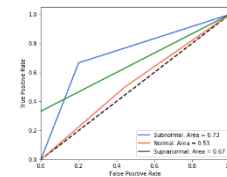


Ensemble_G0_M11_Y2009-2019

MAE: 6.109

Expl. Var.: 0.404

	Precision	Recall	F1-Score
Subnormal	0.73	0.67	0.8
Normal	0.29	0.5	0.2
Supranormal	0.5	0.33	1.0

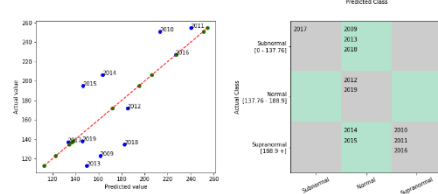
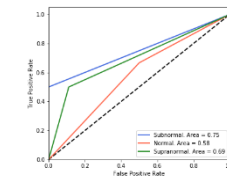


Ensemble_G1_M01_Y2009-2019

MAE: 31.258

Expl. Var.: 0.606

	Precision	Recall	F1-Score
Subnormal	0.67	0.5	1.0
Normal	0.44	0.67	0.33
Supranormal	0.5	0.5	0.5

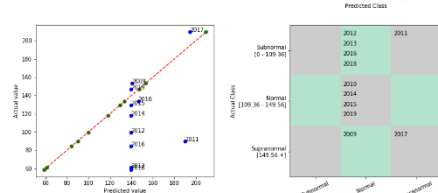
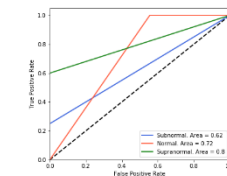


Ensemble_G1_M02_Y2009-2019

MAE: 26.611

Expl. Var.: 0.587

	Precision	Recall	F1-Score
Subnormal	0.4	0.25	1.0
Normal	0.44	1.0	0.29
Supranormal	0.75	0.6	1.0

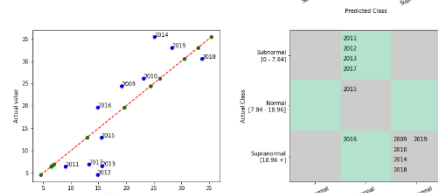
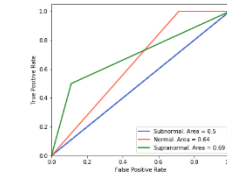


Ensemble_G1_M03_Y2009-2019

MAE: 39.482

Expl. Var.: 0.17

	Precision	Recall	F1-Score
Subnormal	0.0	0.0	0.0
Normal	0.62	1.0	0.44
Supranormal	0.5	0.5	0.5

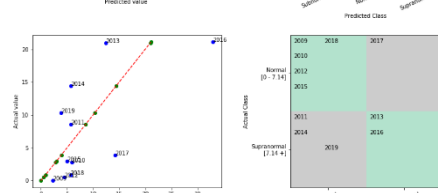
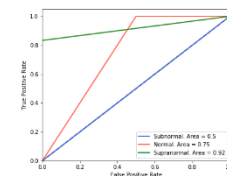


Ensemble_G1_M05_Y2009-2019

MAE: 5.59

Expl. Var.: 0.688

	Precision	Recall	F1-Score
Subnormal	0.0	0.0	0.0
Normal	0.29	1.0	0.17
Supranormal	0.91	0.83	1.0

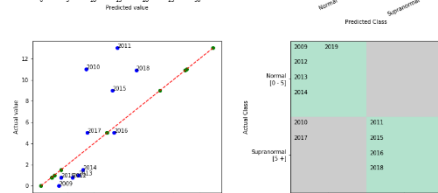
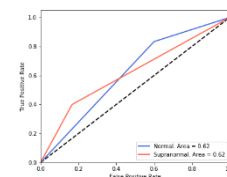


Ensemble_G1_M06_Y2009-2019

MAE: 5.907

Expl. Var.: 0.214

	Precision	Recall	F1-Score
Normal	0.71	0.83	0.62
Supranormal	0.5	0.4	0.67

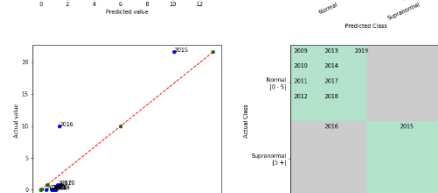
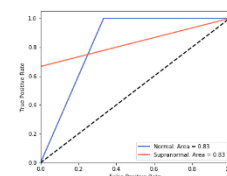


Ensemble_G1_M07_Y2009-2019

MAE: 2.845

Expl. Var.: 0.468

	Precision	Recall	F1-Score
Normal	0.83	1.0	0.71
Supranormal	0.8	0.67	1.0

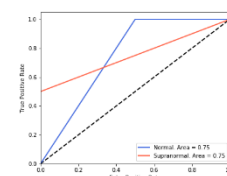


Ensemble_G1_M08_Y2009-2019

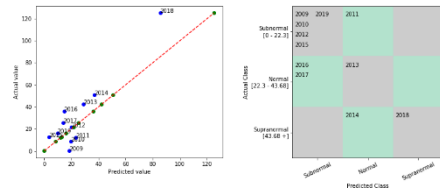
MAE: 2.321

Expl. Var.: 0.781

	Precision	Recall	F1-Score
Normal	0.95	1.0	0.9
Supranormal	0.67	0.5	1.0



Tesis - Juan Ignacio Mazza

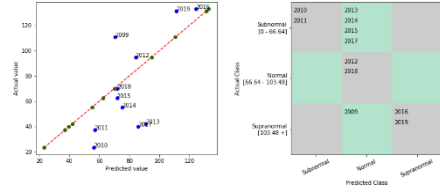
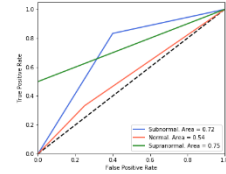


Ensemble_G1_M10_Y2009-2019

MAE: 14.246

Expl. Var.: 0.771

	Precision	Recall	F1-Score
Subnormal	0.77	0.83	0.71
Normal	0.33	0.33	0.33
Supranormal	0.67	0.5	1.0

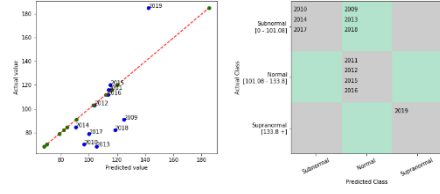
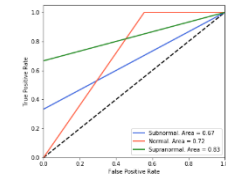


Ensemble_G1_M11_Y2009-2019

MAE: 23.499

Expl. Var.: 0.485

	Precision	Recall	F1-Score
Subnormal	0.5	0.33	1.0
Normal	0.44	1.0	0.29
Supranormal	0.8	0.67	1.0

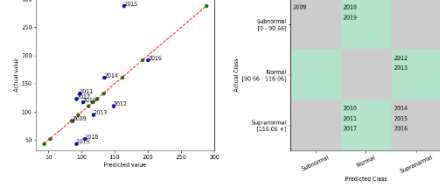
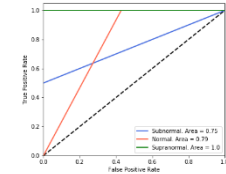


Ensemble_G2_M01_Y2009-2019

MAE: 19.336

Expl. Var.: 0.482

	Precision	Recall	F1-Score
Subnormal	0.67	0.5	1.0
Normal	0.73	1.0	0.57
Supranormal	1.0	1.0	1.0

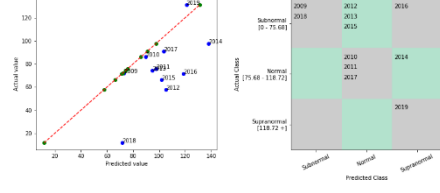
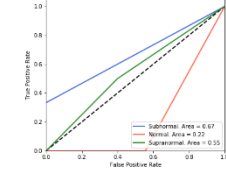


Ensemble_G2_M02_Y2009-2019

MAE: 36.877

Expl. Var.: 0.461

	Precision	Recall	F1-Score
Subnormal	0.5	0.33	1.0
Normal	0.0	0.0	0.0
Supranormal	0.55	0.5	0.6

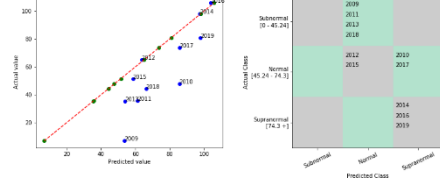
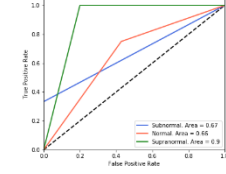


Ensemble_G2_M03_Y2009-2019

MAE: 27.232

Expl. Var.: 0.404

	Precision	Recall	F1-Score
Subnormal	0.5	0.33	1.0
Normal	0.6	0.75	0.5
Supranormal	0.5	1.0	0.33

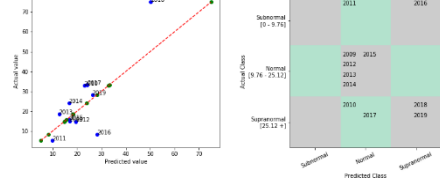
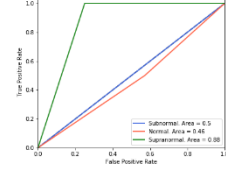


Ensemble_G2_M04_Y2009-2019

MAE: 17.24

Expl. Var.: 0.706

	Precision	Recall	F1-Score
Subnormal	0.0	0.0	0.0
Normal	0.4	0.5	0.33
Supranormal	0.75	1.0	0.6

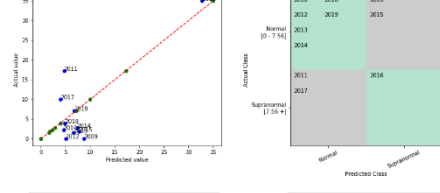
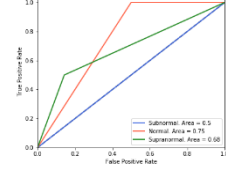


Ensemble_G2_M05_Y2009-2019

MAE: 8.29

Expl. Var.: 0.652

	Precision	Recall	F1-Score
Subnormal	0.0	0.0	0.0
Normal	0.77	1.0	0.62
Supranormal	0.57	0.5	0.67

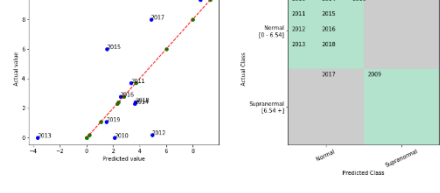
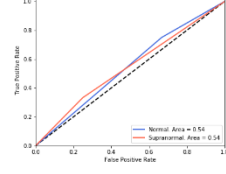


Ensemble_G2_M06_Y2009-2019

MAE: 4.875

Expl. Var.: 0.657

	Precision	Recall	F1-Score
Normal	0.75	0.75	0.75
Supranormal	0.33	0.33	0.33

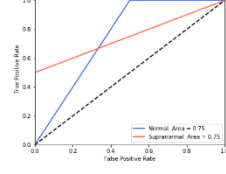


Ensemble_G2_M07_Y2009-2019

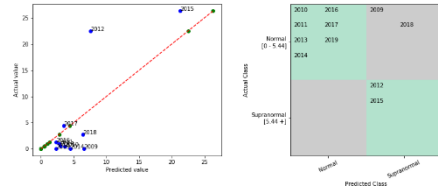
MAE: 2.047

Expl. Var.: 0.288

	Precision	Recall	F1-Score
Normal	0.95	1.0	0.9
Supranormal	0.67	0.5	1.0



Tesis - Juan Ignacio Mazza

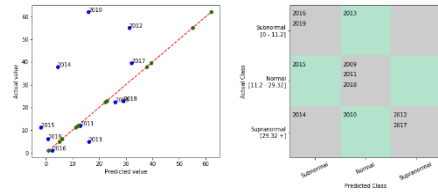
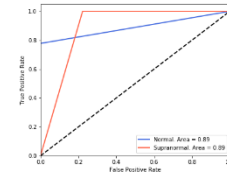


Ensemble_G2_M08_Y2009-2019

MAE: 4.203

Expl. Var.: 0.615

	Precision	Recall	F1-Score
Normal	0.88	0.78	1.0
Supranormal	0.67	1.0	0.5

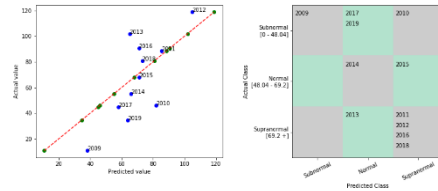
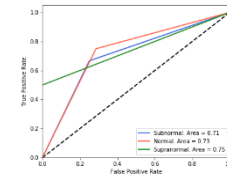


Ensemble_G2_M09_Y2009-2019

MAE: 13.824

Expl. Var.: 0.261

	Precision	Recall	F1-Score
Subnormal	0.57	0.67	0.5
Normal	0.67	0.75	0.6
Supranormal	0.67	0.5	1.0

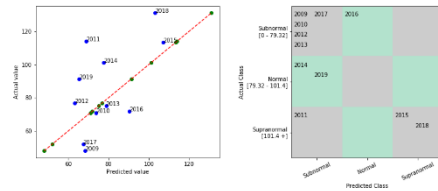
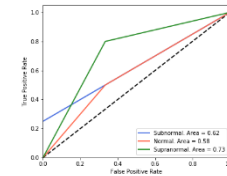


Ensemble_G2_M10_Y2009-2019

MAE: 18.265

Expl. Var.: 0.511

	Precision	Recall	F1-Score
Subnormal	0.4	0.25	1.0
Normal	0.33	0.5	0.25
Supranormal	0.73	0.8	0.67

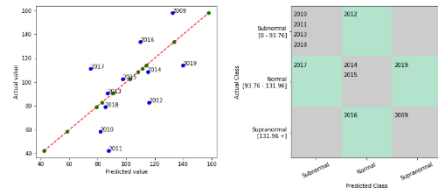
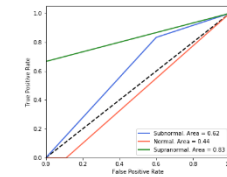


Ensemble_G2_M11_Y2009-2019

MAE: 18.443

Expl. Var.: 0.337

	Precision	Recall	F1-Score
Subnormal	0.71	0.83	0.62
Normal	0.0	0.0	0.0
Supranormal	0.8	0.67	1.0

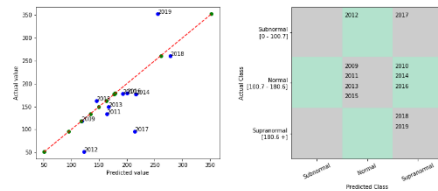
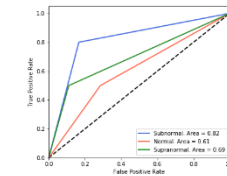


Ensemble_G2_M12_Y2009-2019

MAE: 21.243

Expl. Var.: 0.365

	Precision	Recall	F1-Score
Subnormal	0.8	0.8	0.8
Normal	0.5	0.5	0.5
Supranormal	0.5	0.5	0.5

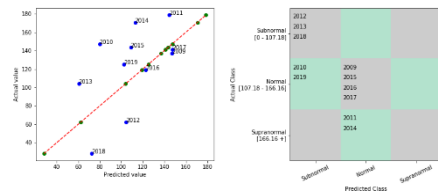
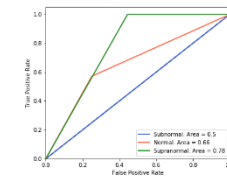


Ensemble_G3_M01_Y2009-2019

MAE: 40.462

Expl. Var.: 0.571

	Precision	Recall	F1-Score
Subnormal	0.0	0.0	0.0
Normal	0.67	0.57	0.8
Supranormal	0.5	1.0	0.33

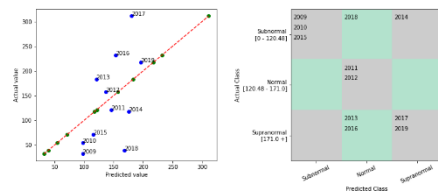
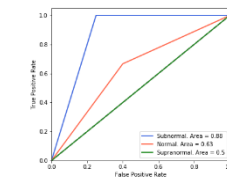


Ensemble_G3_M02_Y2009-2019

MAE: 33.372

Expl. Var.: 0.278

	Precision	Recall	F1-Score
Subnormal	0.73	1.0	0.6
Normal	0.67	0.67	0.67
Supranormal	0.0	0.0	0.0

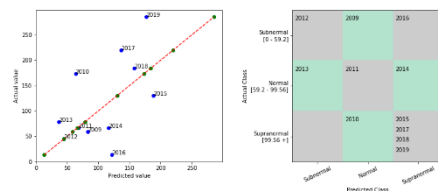
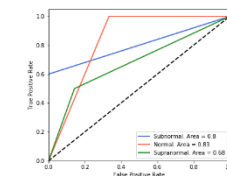


Ensemble_G3_M04_Y2009-2019

MAE: 61.652

Expl. Var.: 0.298

	Precision	Recall	F1-Score
Subnormal	0.75	0.6	1.0
Normal	0.57	1.0	0.4
Supranormal	0.57	0.5	0.67

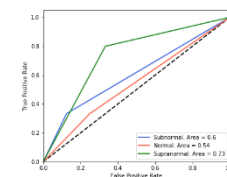


Ensemble_G3_M05_Y2009-2019

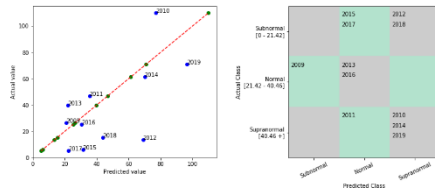
MAE: 55.625

Expl. Var.: 0.312

	Precision	Recall	F1-Score
Subnormal	0.4	0.33	0.5
Normal	0.33	0.33	0.33
Supranormal	0.73	0.8	0.67



Tesis - Juan Ignacio Mazza

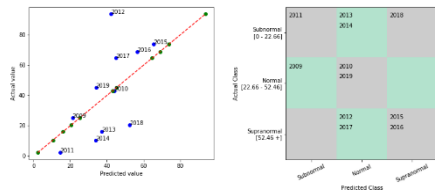
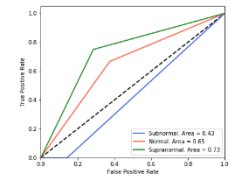


Ensemble_G3_M07_Y2009-2019

MAE: 21.212

Expl. Var.: 0.398

	Precision	Recall	F1-Score
Subnormal	0.0	0.0	0.0
Normal	0.5	0.67	0.4
Supranormal	0.67	0.75	0.6

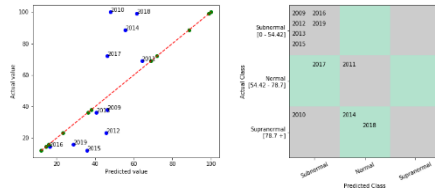
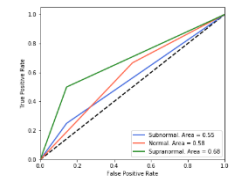


Ensemble_G3_M08_Y2009-2019

MAE: 17.916

Expl. Var.: 0.38

	Precision	Recall	F1-Score
Subnormal	0.33	0.25	0.5
Normal	0.44	0.67	0.33
Supranormal	0.57	0.5	0.67

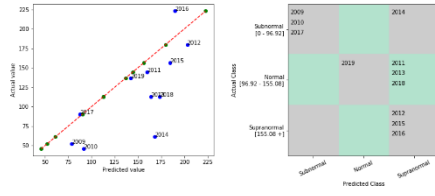
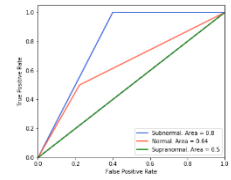


Ensemble_G3_M09_Y2009-2019

MAE: 20.579

Expl. Var.: 0.451

	Precision	Recall	F1-Score
Subnormal	0.86	1.0	0.75
Normal	0.4	0.5	0.33
Supranormal	0.0	0.0	0.0

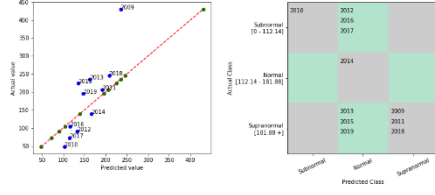
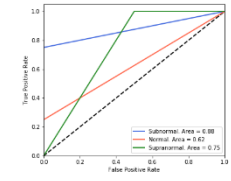


Ensemble_G3_M10_Y2009-2019

MAE: 36.34

Expl. Var.: 0.564

	Precision	Recall	F1-Score
Subnormal	0.86	0.75	1.0
Normal	0.4	0.25	1.0
Supranormal	0.6	1.0	0.43

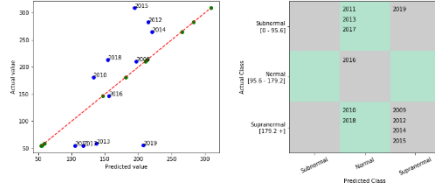
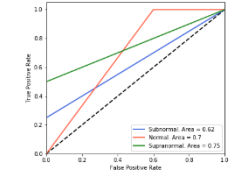


Ensemble_G3_M11_Y2009-2019

MAE: 58.123

Expl. Var.: 0.524

	Precision	Recall	F1-Score
Subnormal	0.4	0.25	1.0
Normal	0.25	1.0	0.14
Supranormal	0.67	0.5	1.0



Ensemble_G3_M12_Y2009-2019

MAE: 63.685

Expl. Var.: 0.368

	Precision	Recall	F1-Score
Subnormal	0.0	0.0	0.0
Normal	0.29	1.0	0.17
Supranormal	0.73	0.67	0.8

