**CONSTRUCCIÓN MODELO PREDICTIVO DE DETERIORO DE PACIENTES DE COVID**

**INSIGHTS DE LA LITERATURA**

**1) Risk prediction models European cohort**

- Uso de SVM, Logística, Boosting, Redes y Bosques aleatorios.

- Conclusiones con datos locales, dudoso generalizar.

- Modelos en diferentes momentos (positivo, admisión a UCI, 12h en UCI, 12h después de UCI).

- Modelos con grupos de características diferentes (básico: edad, sexo, MBI, comorbilidades, temporales, test laboratorio.)

- Bosque aleatorio como modelo escogido (1000 árboles, ajuste hiperparámetros, etc).

- Logística como modelo de referencia.

- Importancia de características a través de mean decrease in impurity (DMI).

- Datos faltantes considerados aleatorios, imputados a través de KNN usando edad y sexo.

- Datos continuos usando medianas comparados con test Mann-Whitney U, categóricos como porcentajes y comparados con chi.

- ROC AUC precisión recall

- Resultados muestran que en el modelo básico de predicción de muerte la edad sale como las variables de más importancia.

- Pre y pos-comorbilidades muestra gran diferencia en el rendimiento del modelo.

- La selección de características mostró que signos vitales y biomarkers son muy importantes para predecir la muerte del paciente.

- No hay pérdida de rendimiento cuando los datos demográficos de los pacientes son similares.

**2) Noticias**

- Madrid, uso de árbol de decisión

- Comorbilidades.

**3) A machine learning-based triage tool for children…**

- Uso de bosque aleatorio con pocas variables.

- Validación cruzada con 5 capas.

**4) Tesis**

- Pruebas con varias técnicas.

**5) A predictive model and risk factors for case fatality of covid-19**

- Un primer modelo con variables demográficas y comorbilidades, un segundo modelo incluyendo saturación de oxígeno, y un tercer modelo incluyendo indicadores de análisis de sangre.

- Importancia de variables con método Shapley Additive Explanation.

- Clasificación por niveles de riesgo.

**PREPROCESAMIENTO DE LOS DATOS**

- Estratificación no necesaria para predicción de muertes con mismo grupo de infectados (es probable que sí)

- Datos faltantes.

Imbalanceado.

- Codificación de variables categóricas.

- Codificación de etiquetas manual.

- one hot encoding (posible problema de dimensionalidad)

- Codificación binaria por límites.

- Scikit learn para hacerle codificación a la variable objetivo.

- Determinación de pesos de acuerdo con criterio experto.

- Reemplazo con pesos (todas las variables con peso).

- Ponderar (que los pesos sumen 1)

**ANÁLISIS DE LOS DATOS**

- Distribuciones

- Prueba Mann – Whitney U y Chi cuadrado.

**MODELO BASE**

* Construcción de regresión logística.
  + Múltiples regresiones como en los paper
    - Datos demográficos, comorbilidades, test de laboratorio, etc.

**MODELO CON RANDOM FOREST**

¿Por qué Random Forest?

Este modelo es el intermedio perfecto entre interpretabilidad y complejidad, es preciso y no es difícil de explicar.

Otro motivo para elegirlo es por su inherente capacidad para detectar efectos de interacción entre las variables, en contraste a la regresión donde se tienen que especificar

Variables demográficas

Edad, sexo, comorbilidad, síntomas (oximetría), uci o no uci

Piso 13 del edificio principal como investigador, doctora Daniela Vega, Unidad de investigación, 2pm.

Reforzar datos clínicos directamente del sistema de monitoreo, parte tecnológica

**PREPROCESAMIENTO**

Edad  
Sexo