# PRUEBA TECNICA CIENTIFICO DE DATOS

**Nombre del Candidato: Julian Andrés Quimbayo Castro**

**Tiempo de la prueba: 3 horas**

La siguiente prueba se debe desarrollar en el programa, funciones y paquetes de su preferencia. Se debe presentar un informe Gerencial y técnico con los resultados y el proceso desarrollado.

1. **Modelo de Fuga**
   1. **Descripción del reto analítico**

Colsubsidio desea establecer la probabilidad que un cliente en una de sus líneas de negocio (Crédito) se vaya o no. Para esto se tiene un histórico de datos de los clientes que han solicitado un tipo de retiro y los que no lo han expresado.

Teniendo en cuenta esta información se entregan las siguientes bases de datos y archivos:

* **diccionario\_datos ->** Contiene una breve descripción de los campos en las BD.
* **train ->** BD con todas las variables del negocio y la variable target.
* **test ->** BD para pobrar los resultados del modelo.
* **train\_test\_demograficas** -> BD con variables sociodemográficas
* **train\_test\_subsidios ->** BD con variables de subsidios

1. **Modelo de Segmentación**

**2.1 Descripción del reto analítico**

Colsubsidio desea segmentar y perfilar sus clientes de Crédito con el objetivo de ofrecerles a través de campañas de marketing tarjetas de crédito y seguros. Para este ejercicio, utilice las fuentes de datos provistas para el primer modelo y construya un modelo que permita segmentar los clientes. Mencione las métricas que tuvo en cuenta para evaluar la confiabilidad de su modelo.

**3. Describa brevemente los siguientes conceptos y mencione los escenarios de aplicación para cada uno de ellos:**

|  |  |  |
| --- | --- | --- |
| **Concepto** | **Descripción** | **Escenario de Aplicación** |
| Sistema de recomendación | Un sistema de recomendación permite hiperpersonalizar el contenido a usuarios con base en intereses y contenidos factibles en particular. Este proceso se realiza con base en gustos, preferencias recopiladas por medio de la interacción con el usuario en un contexto abierto. | Un escenario importante de aplicación lo vemos de manera cotidiana en las plataformas de contenidos como son Netflix, amazon prime, Spotify entre otros que toman los gustos de los usuarios se analizan por medio de algoritmos de machine learning y sugieren música, películas o series acorde a dicha interacción. |
| Sesgo estadístico | El sesgo examina la diferencia entre la medición promedio observada y un valor como referencia esto permite determinar la exactitud del sistema de medición. | Un Escenario claro es para las variables aleatorias numéricas existen las pruebas de normalidad que sirven por medio del uso de pruebas como shapiro-wilk y kolmogorov-smirnov la aceptación o no de una hipótesis a nivel de investigación. En educación se utiliza mucho para determinar si la pregunta de investigación y su hipótesis es aceptable o no. |
| Análisis Multivariado | El análisis multivariado es el uso de técnicas descriptivas y explicativos de las diferentes observaciones dentro de los datos con el fin de comenzar a determinar insights significativos que permitan determinar el enfoque del estudio a realizar. | Todos los escenarios para realizar un análisis descriptivo exploratorio (EDA), son relevantes ya que siempre se debe conocer la data y el objetivo del análisis. Esto permite entender las fuentes, naturaleza y alcance de la data frente a los objetivos corporativos y modelos de negocio. |
| Aprendizaje supervisado | El aprendizaje supervisado permite responder preguntas concretas del modelo de negocio que se esté analizando, es decir, indagar si con la data obtenida es posible enfocar estrategias de mejora de negocio o no y como poder generar ventajas competitivas. | Actualmente, se utiliza mucho para determinar que características o variables de diferentes procesos son las que se deben tener en cuenta o no para mejorar el servicio al cliente, u ofrecer productos y servicios nuevos que a simple vista no se tendrían en cuenta. Poder estimar el valor de una casa con base en diferentes características o saber si un estudiante pierde o no una asignatura con base en su comportamiento social. |
| Aprendizaje No-supervisado | El aprendizaje No-Supervisado consiste en sumergirse en una gran cantidad de datos y encontrar insights sin tener en cuenta una pregunta en concreto. Se usa mucho en temas de clustering o generación de grupos de interés o reducción de dimensionalidad de un set de datos muy grande. | Un caso en concreto es la generación de nichos de mercado para marketing con técnicas de clustering utilizando intereses en común, esto posibilita nuevos espacios para impactar con campañas de marketing eficientes y siendo conscientes de los gustos particulares de cada individuo. |
| Análisis de conglomerados | Este análisis permite determinar un determinado número de grupos de variables por medio de la similitud entre observaciones, se usan dos tipos uno jerárquico y otro por medio del valor de una constante K que determina un solo método de agrupación y la cantidad de grupos a crear. Internamente utiliza la distancia entre una observación y otra y así por medio del valor de K determinar a que grupo pertenece según ese valor. | Un escenario importante es la generación de grupos en marketing digital para así enfocar campañas de atracción de leads, es un uso muy constante y en detección de anomalías como en fraudes bancarios. |
| Random Forest | Es un algoritmo de aprendizaje supervisado orientado por la generación de arboles de clasificación y regresión. Utiliza dos técnicas para el entrenamiento que son Bagging y Boosting, la primera permite generar una muestra aleatorio generando múltiples arboles teniendo en cuenta la variable target a clasificar o predecir y dependiendo del árbol que mayor peso tenga se selecciona las mejores opciones; la segunda, realiza el mismo proceso aleatorio pero de manera secuencial, esto garantiza lidiar de mejor manera con las variables que tengan desbalance y establecer de manera mas acertada cuales son las variables mas o menos importantes para clasificar o predecir. | Si se desea saber que variables de un proceso en particular son las relevantes o menos relevantes a tener en cuenta esto permite reconfigurar estrategias comerciales, marketing, ventas, logística entre otras. |
| Cross-Validation | La validación cruzada permite determinar por medio de pequeñas porciones de la data para testear el modelo en un ambiente de producción. | Todos los modelos sobre todo de predicción se hacen necesario y/o obligatorio realizar esta validación para simular un ambiente de producción y verificar por medio de otras métricas la eficiencia, especificidad, sensibilidad y margen de error del modelo. |
| Test de silhouette | Es un método de interpretación y validación de la coherencia a la hora de realizar agrupaciones por medio de técnicas como K-means. Dicha técnica esta enmarcada entre -1 y +1, entre más este el valor alto indica mayor cohesión, si por el contrario el valor es menor indica que hay muchos cúmulos es decir una baja cohesión. | Todos los algoritmos usados para procesos de clustering deben realizar esta prueba que sirve como proceso de calidad en el ciclo de vida de Machine Learning. |
| Curva ROC | Esta curva permite como métrica identificar la relación de un clasificador entre la sensibilidad y la especificidad de un modelo. Es decir, la relación entre los casos de éxito y las falsas alarmas a la hora de detectar casos en producción. Es muy útil de manera gráfica y numérica para saber que tan fiable es un modelo cuando se necesita detectar casos positivos o negativos. | Se utiliza mucho en problemas clasificación binarios, como saber si un paciente tiene cáncer o no, saber porque se escoge un vino de otro entre otros. |