Guía de próximos pasos:

*Comentarios del tutor en reunión del 20-06-24*

**Revisar y Familiarizarse con el Dataset:**

* Profundizar en la comprensión de cada variable, especialmente aquellas que puedan tener un impacto significativo en la predicción de enfermedades cardiovasculares. Examinar las relaciones entre variables utilizando gráficos de correlación.

**Análisis de Clustering:**

* Considerar realizar un análisis de clustering para identificar patrones o grupos naturales en los datos. Esto puede ayudar a descubrir subgrupos de pacientes con perfiles de riesgo similares, lo cual es útil para personalizar las recomendaciones de prevención y tratamiento.

**Feature Engineering:**

* Evaluar qué nuevas variables podrían ser creadas a partir de las existentes para mejorar el modelo. Por ejemplo, se podría combinar variables relacionadas con el estilo de vida o crear interacciones entre la edad y los biomarcadores.

**Importancia de las Características con Random Forest:**

* Implementar un modelo de Random Forest para evaluar la importancia de las características. Este análisis permitiría identificar qué variables tienen el mayor impacto en la predicción de enfermedades cardiovasculares, lo que puede influir en las decisiones clínicas y en la priorización de intervenciones.

**Validación del Desempeño del Modelo:**

* Utilizar técnicas robustas de validación para evaluar el desempeño de los modelos. La validación cruzada es esencial para asegurar que tus modelos son estables y generalizables.

**Ideas adicionales para el feature engineering a considerar:**

**Interacciones entre Variables:**

Es posible crear nuevas variables que sean el producto de dos o más variables que se sospeche interactúan entre sí.

**Variables Polinomiales:**

Considerar generar términos polinomiales (cuadráticos, cúbicos) de variables que creas que tienen relaciones no lineales con la variable objetivo.

**Variables Categóricas Agrupadas:**

Si hay variables categóricas con muchas categorías, se puede intentar agruparlas en categorías más grandes que tengan sentido clínico o estadístico.

**Indicadores de Riesgo Basados en Umbrales:**

Crear variables indicadoras (0/1) basadas en umbrales clínicamente significativos.

**Desagregar Variables Complejas:**

Si existen variables que son compuestos de varios factores (como un índice de salud general), considerar desglosarlos en componentes individuales si los datos lo permiten.

**Índices Compuestos:**

Desarrollar índices compuestos que combinen múltiples medidas en un solo indicador.

Antes de avanzar a la etapa de modelado es posible agregar algunas acciones adicionales para enriquecer aún más el análisis:

1. **Análisis de Correlación Completo:**
   * Realizar un análisis de correlación más exhaustivo, no solo entre las variables independientes y la variable dependiente, sino también entre las variables independientes mismas. Esto puede ayudar a identificar multicolinealidad o relaciones interesantes que podrían influir en el modelado.
   * Utilizar mapas de calor para visualizar estas correlaciones.
2. **Técnicas Avanzadas de Reducción de Dimensionalidad:**
   * Aplicar técnicas como Análisis de Componentes Principales (PCA) para reducir la dimensionalidad del dataset. Esto puede ser especialmente útil si hay varias variables y se desea simplificar el modelo sin perder información crítica.
   * Explorar cómo la reducción de dimensiones afecta la capacidad predictiva del modelo.
3. **Extracción y Selección de Características Avanzadas:**
   * Implementar métodos automáticos de selección de características, como la eliminación recursiva de características (RFE) o métodos basados en modelos (como los coeficientes de un modelo de regresión Lasso) para identificar las variables más importantes.
   * Considerar la creación de características sintéticas que puedan capturar la interacción entre variables de maneras no lineales que no se haya considerado anteriormente.
4. **Evaluación de Outliers y Anomalías:**
   * Realizar un análisis detallado de outliers para entender cómo afectan al modelo. Los outliers pueden tener un impacto significativo en los modelos de regresión y pueden necesitar ser tratados o excluidos.
   * Considerar métodos para manejar outliers, como transformaciones logarítmicas o truncamiento.
5. **Validación de Hipótesis:**
   * Antes de proceder al modelado, validar cualquier hipótesis que se haya formulado durante la fase de exploración de datos. Esto podría incluir pruebas estadísticas específicas para verificar suposiciones sobre las distribuciones de los datos o la relación entre variables.
6. **Optimización de Hiperparámetros para Clustering:**
   * Asegurarse de que los parámetros utilizados en las técnicas de clustering, como el número de clusters en K-means, sean óptimos. Es posible usar métodos como el método del codo o la silueta para determinar el número adecuado de clusters.
7. **Documentación Detallada:**
   * Documentar meticulosamente todos los pasos realizados hasta ahora, incluyendo la justificación de decisiones importantes como la elección de técnicas de feature engineering y los métodos de análisis.

\*Todo esto basado en lo aprendido durante el módulo 7 del máster.