Propuesta de Modelo de Segmentación para Identificación de Riesgos de Salud Laboral

1. Introducción

Empathy Salud, una empresa líder en servicios de medicina ocupacional preventiva, busca mejorar su capacidad analítica y predictiva mediante el uso de datos médicos anonimizados. Este documento presenta una propuesta de modelo de segmentación basado en datos para identificar riesgos de salud laboral, incorporando técnicas de inteligencia artificial generativa para ofrecer recomendaciones personalizadas y explicaciones detalladas.

El análisis de riesgos de salud laboral es fundamental para la prevención en entornos ocupacionales. La segmentación de trabajadores en grupos de riesgo permite priorizar intervenciones efectivas, lo que reduce costos médicos y mejora el bienestar de los empleados. La incorporación de IA generativa agrega valor al proporcionar recomendaciones más personalizadas y comprensibles, facilitando la toma de decisiones por parte de empleadores y profesionales de la salud.

2. Lógica y Respaldo Metodológico

El modelo propuesto sigue una metodología estructurada en las siguientes etapas:

2.1 Preprocesamiento de Datos

- **Limpieza y Normalización:** Se eliminan valores atípicos, se imputan datos faltantes y se normalizan variables numéricas.
- **Codificación de Variables:** Transformación de datos categóricos en variables numéricas mediante técnicas como one-hot encoding o embeddings.
- **Reducción de Dimensionalidad:** Aplicación de PCA o t-SNE para mejorar la interpretabilidad y reducir el ruido en los datos.

2.2 Segmentación de Trabajadores

- Se utilizarán algoritmos de clustering como **K-Means**, **DBSCAN** o **Gaussian Mixture Models (GMM)** para agrupar trabajadores según factores de riesgo como:
    - Índice de Masa Corporal (IMC)
    - Presión arterial
    - Niveles de colesterol

    - Historial de ausentismo

- Se evaluarán diferentes métricas de calidad de clustering (Silhouette Score, Davies-Bouldin Index) para seleccionar el modelo óptimo.

2.3 Integración de IA Generativa

- Se implementará un modelo basado en **GPT-Neo** para:
    - Generar recomendaciones clínicas personalizadas según el segmento de riesgo.
    - Producir explicaciones claras en lenguaje natural para mejorar la comprensión de los hallazgos.
- Se entrenará el modelo con un corpus de literatura médica y guías clínicas relevantes para garantizar la precisión de las recomendaciones.

El respaldo metodológico se fundamenta en guías clínicas internacionales y literatura especializada en medicina ocupacional, asegurando que las decisiones y recomendaciones sean válidas y basadas en evidencia.

3. Implementación Técnica

El pipeline de datos incluirá las siguientes fases:

3.1 Extracción y Transformación de Datos

- Uso de **Python** y librerías como **pandas**, **NumPy** y **Scikit-learn** para la manipulación de datos.
- Integración con bases de datos mediante **SQL** y procesamiento en entornos de Big Data con **PySpark**.

3.2 Segmentación y Análisis

- Aplicación de algoritmos de clustering sobre datos anonimizados.
- Evaluación de patrones y correlaciones utilizando técnicas de visualización con **Seaborn** y **Matplotlib**.

3.3 Generación de Insights con IA Generativa

- Implementación de **GPT-Neo** para la generación de informes personalizados.
- Creación de dashboards interactivos con **Power BI** o **Streamlit** para facilitar la interpretación de los resultados.

4. Aplicabilidad Práctica

Este enfoque tiene un alto potencial de aplicabilidad en entornos empresariales y de salud ocupacional. Sus principales beneficios incluyen:

- **Reducción de costos asociados a enfermedades ocupacionales** mediante la identificación temprana de riesgos y la aplicación de medidas preventivas.
- **Mejora en la comunicación con los trabajadores**, proporcionando explicaciones claras y personalizadas sobre su estado de salud y recomendaciones preventivas.
- **Incremento en la precisión de las recomendaciones**, gracias a la combinación de segmentación avanzada y modelos generativos de IA.
- **Facilitación en la toma de decisiones** para profesionales de la salud ocupacional y recursos humanos, optimizando la asignación de recursos preventivos.

5. Conclusión

La propuesta presentada combina técnicas avanzadas de análisis de datos y generación de texto con IA para ofrecer soluciones prácticas y efectivas en la identificación y gestión de riesgos de salud laboral. Al integrar un enfoque basado en datos con herramientas de interpretabilidad, se busca maximizar el impacto de las estrategias de medicina ocupacional preventiva.

6. Referencias

- Organización Mundial de la Salud. (2021). Guías sobre salud ocupacional.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Vaswani, A., et al. (2017). *Attention is all you need*. Advances in Neural Information Processing Systems.