**TITULO PROYECTO:** Depression Risk Analysis

**INTEGRANTES:** Luis Alejandro Londoño Martínez, Simón Correa Marín

1. **ENTENDIMIENTO DEL NEGOCIO**
   1. DESCRIPCIÓN DEL NEGOCIO

El **proyecto Depression Risk Analysis** se centra en analizar los factores de riesgo que pueden influir en la probabilidad de desarrollar depresión en adultos. Su objetivo principal es identificar patrones y correlaciones en las respuestas de una encuesta de salud mental para predecir el riesgo de depresión, basándose en datos de estilo de vida y demográficos de personas entre 18 y 60 años. A través de este análisis, se espera aportar a la comprensión de cómo variables como la satisfacción laboral, horas de estudio o trabajo, antecedentes familiares y otros factores de la vida cotidiana se relacionan con el riesgo de sufrir depresión.

Este proyecto se sitúa en el contexto de la **investigación en salud mental,** específicamente en la identificación de factores clave que afectan el bienestar psicológico en un entorno no clínico. La información obtenida se utilizará para desarrollar modelos predictivos de aprendizaje supervisado que podrían emplearse en la formulación de estrategias preventivas y en la toma de decisiones para intervenciones tempranas en el ámbito de la salud pública.

* 1. DESCRIPCIÓN DEL PROBLEMA

El problema específico que busca resolver este proyecto es identificar los factores cotidianos y demográficos que se asocian con el riesgo de depresión en adultos. A falta de evaluaciones clínicas y diagnósticos profesionales en este contexto, el desafío radica en emplear un modelo predictivo basado en datos de estilo de vida y satisfacción personal para predecir con precisión el riesgo de depresión en un grupo poblacional diverso. Esta herramienta permitirá entender mejor cómo ciertos factores pueden influir en la salud mental y proporcionar una base para la intervención temprana y la toma de decisiones informadas en materia de bienestar.

El problema que se busca abordar es la **falta de herramientas predictivas para evaluar el riesgo de depresión** en adultos con base en factores no clínicos y de estilo de vida.

* 1. OBJETIVOS DE LA MINERÍA
* Identificar los factores clave asociados al riesgo de depresión en adultos mediante un análisis exploratorio y predictivo de los datos recolectados.
* Construir modelos de aprendizaje supervisado para la clasificación del riesgo de depresión. Se emplearán cinco algoritmos de clasificación:
  + Máquinas de soporte vectorial para clasificación (SVM)
  + Red Neuronal para clasificación (ANN)
  + Árboles de decisión
  + K-Nearest Neighbors (KNN)
  + Regresión Logística

Estos modelos se entrenarán sobre un conjunto de datos balanceado al 70%.

* Aplicar cuatro métodos de ensamble para mejorar la robustez y precisión de los modelos predictivos:
  + Random Forest, que combina múltiples árboles de decisión de forma aleatoria.
  + XGBoost y CatBoost, técnicas avanzadas de boosting que ajustan iterativamente los modelos para minimizar los errores residuales.
  + **Voting Hard**, un ensamble que combina múltiples modelos de clasificación mediante un sistema de votación. En este caso, se utiliza un voto mayoritario, donde la clase final se determina según la clase que reciba el mayor número de votos de todos los modelos individuales.
* Comparar y evaluar el rendimiento de cada modelo utilizando al menos cuatro métricas de calidad, como precisión, recall, F1-score y curva ROC. Estas métricas permitirán identificar los modelos más eficientes y adecuados para la predicción del riesgo de depresión.
* Seleccionar los tres mejores modelos utilizando un análisis estadístico de diferencia significativa (ANOVA y prueba de Tukey) para garantizar que la elección de los modelos tenga una base estadísticamente válida.
* Optimizar los tres modelos seleccionados mediante hiperparametrización usando GridSearch y métodos avanzados como BayesSearchCV (optimización bayesiana) y **GASearchCV** (algoritmos genéticos) para maximizar el desempeño predictivo de cada modelo.
* **Desplegar el modelo final** mediante un Pipeline de preparación de datos, integrándolo en una interfaz gráfica desarrollada con **Streamlit.** Esta interfaz permitirá al usuario cargar datos, ejecutar el modelo y visualizar las predicciones de riesgo, proporcionando una herramienta funcional y accesible para la toma de decisiones en salud mental y bienestar psicológico. El despliegue se realizará con **LocalTunnel** para generar un enlace público que facilite el acceso remoto a la aplicación, permitiendo que otros usuarios interactúen con la herramienta de forma sencilla y segura.
  1. DISEÑO DE SOLUCIÓN

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Problema | Tipo de Análisis | Tipo de Aprendizaje | Tarea Analítica | Requerimientos en los Datos | Métodos | Evaluación | Calidad Esperada |
| Construir modelos de clasificación para predecir el riesgo de depresión e identificar los factores asociados al riesgo. | Predictivo | Supervisado | Análisis exploratorio y clasificación del riesgo | Datos de calidad con variables demográficas, sociales y laborales balanceados al 70% para evitar sesgos | Modelos: SVM, ANN, Árbol de decisión, KNN y Regresión Logística | Matriz de confusión, Precisión, recall, F1-score, curva ROC | Modelos de clasificación con precisión y sensibilidad superiores al 80%. |
| Mejorar la precisión de los modelos utilizando técnicas de ensamble | Predictivo | Supervisado | Ensamble para precisión y estabilidad | Datos limpios y preprocesados; balanceo previo al 70%; partición en conjuntos de entrenamiento y prueba | Ensambles: Random Forest, XGBoost, CatBoost, Voting Hard (con voto mayoritario) | Precisión, recall, F1-score, curva ROC | Modelos ensamblados con desempeño superior a modelos individuales. Aumento en precisión y robustez, con mejora del 5-10% en métricas sobre modelos individuales. |
| Seleccionar los mejores modelos mediante un análisis estadístico de significancia | No aplica | No aplica | Comparación y selección de modelos | Conjunto de datos de entrenamiento y validación, asegurando balanceo previo | ANOVA y prueba de Tukey | Diferencia estadística en rendimiento entre modelos | Selección de modelos basada en significancia estadística con una confiabilidad mínima del 95%. |
| Optimizar el rendimiento de los modelos seleccionados a través de hiperparametrización | No aplica | No aplica | Optimización de hiperparámetros | Conjunto de datos de validación balanceado para GridSearch y optimización avanzada de hiperparámetros | Hiperparametrización con GridSearch, BayesSearchCV y GASearchCV para modelos seleccionados | Matriz de confusión, Precisión, recall, F1-score, curva ROC tras optimización | Mejora en el rendimiento de los modelos seleccionados en al menos un 5% en comparación con los resultados iniciales. |
| Desplegar el modelo final en una interfaz gráfica para predicción del riesgo de depresión | No aplica | No aplica | Despliegue mediante una interfaz gráfica | Pipeline de datos con preprocesamiento completo, normalización, codificación y preparación de datos para despliegue | Implementación en Streamlit, con despliegue público mediante LocalTunnel | No aplica | Interfaz funcional y amigable para el usuario |

* 1. RECURSOS PARA CREACIÓN DEL MODELO Y PARA DESPLIEGUE

|  |  |
| --- | --- |
| Aspecto | Detalles |
| Entorno de Desarrollo | La creación del modelo se llevará a cabo de forma **local** utilizando **Jupyter Notebook** en **Visual Studio Code**. |
| Lenguaje de Programación | Se empleará **Python** tanto para la creación y entrenamiento del modelo como para el despliegue en Streamlit. |
| Librerías y Herramientas | - **Limpieza de datos**: pandas, numpy, sklearn  - **Entrenamiento y evaluación de modelos**: scikit-learn - **Optimización y búsqueda de hiperparámetros**: GridSearchCV, BayesSearchCV, GASearchCV  - **Visualización de datos**: matplotlib, seaborn  - **Despliegue**: Streamlit, LocalTunnel |
| Repositorio y Control de Versiones | Los cambios del proyecto se subirán a un repositorio remoto en **GitHub**, permitiendo un control de versiones adecuado y colaboración si es necesario. |
| Despliegue del Modelo | - El modelo se desplegará en una aplicación de **Streamlit**, donde los usuarios podrán cargar datos, ejecutar el modelo y ver los resultados en una interfaz gráfica.  - Se utilizará **LocalTunnel** para generar un enlace público y permitir el acceso remoto a la aplicación, facilitando el uso del modelo en otros dispositivos o ubicaciones sin necesidad de configuración compleja. |
| Interacción Modelo-Interfaz | La interfaz de Streamlit cargará el modelo desde un archivo preentrenado almacenado localmente. Streamlit ejecutará el modelo en tiempo real y mostrará los resultados al usuario, asegurando una comunicación directa entre el modelo y la interfaz. |
| Licencias y Requisitos Legales | - **Licencia de Python y bibliotecas**: Python es de código abierto, y las bibliotecas seleccionadas tienen licencias que permiten el uso y modificación libre. - **Visual Studio Code**: de código abierto, con licencias de uso gratuito para desarrolladores.  - **Streamlit y LocalTunnel**: también de código abierto y gratuitos. |
| IDE | **Visual Studio Code** para la creación y pruebas locales del modelo en Jupyter Notebooks. |

1. **ENTENDIMIENTO DE LOS DATOS**
   1. CICLO DE LOS DATOS

**Generación de los datos**  
Los datos se generan a partir de una encuesta anónima que fue diseñada específicamente para evaluar factores demográficos, sociales y de estilo de vida relacionados con el riesgo de depresión en adultos. La encuesta fue distribuida en varias ciudades y capturó información autodeclarada de personas de entre 18 y 60 años sobre factores como la satisfacción laboral, horas de trabajo/estudio, antecedentes familiares, entre otros.

**Almacenamiento de los datos**  
Actualmente, los datos recolectados están almacenados en archivos CSV que han sido descargados de la plataforma Kaggle. Estos archivos se gestionan de manera local, pero se almacenan en un repositorio remoto en GitHub para facilitar el control de versiones y permitir la colaboración entre los integrantes del equipo.

**Modificación de los datos**  
Dado que se trabaja con archivos descargados, la manipulación de datos se hace por los miembros del equipo de análisis de datos. Las operaciones de preprocesamiento, limpieza y transformación de datos se realizarán mediante scripts en Python.

**Periodicidad de los datos y reentrenamiento del modelo.**  
Los datos de la encuesta fueron recolectados en un solo periodo (enero a junio de 2023), lo que significa que el conjunto de datos es estático y no recibirá actualizaciones periódicas. En caso de futuras expansiones del proyecto o de la encuesta, se podrían establecer nuevas fases de recolección de datos. De ser así, el modelo podría ser reentrenado cada 2-3 años si se recopilan datos adicionales que reflejen cambios en los factores de riesgo para la depresión.

* 1. DICCIONARIO DE DATOS

Tiene que existir (en teoría). El diccionario de datos de la base de datos describe la variable, el tipo de la variable y su descripción.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Tipo** | **Descripción** |
|  |  |  |

* 1. REGLAS DE CALIDAD

No se construyen a partir de los datos.

|  |  |
| --- | --- |
| **Variable** | **Regla de calidad** |
| Edad | 18 - 60 |

Cualquier valor fuera del rango de datos, es un error. Estas reglas las aporta un experto en TI del negocio. Reglas de calidad por cada variable. Servirá en un futuro para identificar atípicos.

1. **PREPARACIÓN DE DATOS (Estadística)**
   1. INTEGRACIÓN

Se crea la sabana de datos, unir tablas (join).

* 1. SELECCIÓN DE VARIABLES

Borrar variables irrelevantes, las que no sirven para el modelo o que hagan parte de una regulación de privacidad de datos (nombres, direcciones, cédulas, teléfonos).

* 1. DESCRIPCIÓN ESTADÍSTICA

Medidas estadísticas, gráficos, etc.

* 1. LIMPIEZA DE ATÍPICOS

Cualquier valor por fuera del rango expuesto en las reglas de calidad es considerado atípico.

* 1. LIMPIEZA DE NULOS

No siempre se puede hacer imputación. Definir si borro o imputo (inventar datos).

* 1. CREACIÓN DE NUEVAS VARIABLES

Creación de nuevas variables a partir de datos libres.

* 1. ANÁLISIS DE CORRELACIONES PARA REDUNDANCIA

Se usa para eliminar variables que no me sirven porque son redundantes.

* 1. ANÁLISIS DE CORRELACIONES PARA IRRELEVANCIA (PREDICCIONES)

Cuando una variable definitivamente no me sirve para nada.

* 1. REDUCCIÓN DE DIMENSIÓN (OPCIONAL EN PREDICCIONES)

Análisis de componentes principales (PCA), se asumen riesgos como perdida de información al reducir variables. Puedo perder el 0.25 de la varianza original de los datos.

* 1. BALANCEO (CLASIFICACIÓN)

Cuando la variable objetivo está muy desbalanceada se puede optar por hacer balanceo sobre muestreo. Muy riesgoso también porque evalúas mediante datos ficticios.

* 1. TRANSFORMACIONES

Tener en cuenta si debo crear dummies o no, es decir, tener en cuenta con que estoy trabajando.

1. **MODELAMIENTO, EVALUACIÓN E INTERPRETACIÓN**
   1. CONFIGURACIÓN MÉTODOS DE MACHINE LEARNING

Se aplican los métodos que se especificaron en el diseño de solución, y se configuran.

* 1. ANALISIS DE MEDIDAS DE CALIDAD

Análisis de las medidas de calidad definidas previamente.

* 1. SELECCIÓN DEL MEJOR MODELO

Además de la medida de calidad se deben tener en cuenta otros aspectos como la complejidad computacional. Se usan métodos como la ANOVA o la prueba de Tuckey. Si no hay diferencia significativa en los modelos se parte de la comparación a partir de la complejidad computacional.

1. **DESPLIEGUE**
   1. PREDICCIÓN DE DATOS FUTUROS

Despliegue por interfaz gráfica, app, nube, web, etc.

* 1. CRONOGRAMA DE MANTENIMIENTO

Haya o no errores en el modelo, se hace un mantenimiento cada cierto periodo de tiempo definidos en las primeras etapas.