STUDENT DEPRESSION PREDICTION

DANIEL LOEVETSKI 209059120
YUVAL HOFFMAN 206202533

INTRODUCTION

Our project explores the critical issue of mental illness and specifically depression among students. This subject is particularly sensitive and significant as it impacts students' well-being and academic success directly. By leveraging Big Data to gain insights from extensive datasets we aim to analyze a vast array of behavioral and demographic data from diverse sources. This approach underscores the potential of Big Data to advance mental health interventions and support systems within educational environments.

THE PROBLEM'S DOMAIN

Our project focuses on identifying university students at higher risk for depression by recognizing specific risk factors and indicators. By analyzing behavior, academic performance, and personal circumstances, we aim for proactive identification that allows for early intervention. This approach is essential for managing mental health risks, ultimately fostering a supportive educational environment that enhances student welfare and success.

DATASET

The dataset for our project was sourced from Kaggle.

It is structured as a CSV file that contains 18 columns and 27,900 rows representing students from India.

This dataset includes behavioral, socio-economical, and demographical parameters that help identify potential depression among students, such as academic pressure, academic pressure, city, family history of mental illness, and more.

These various factors provide a comprehensive foundation for analyzing the different dimensions

id	Gender	Age	City	Profession	Academic Pressure	Work Pressure	CGPA	Study Satisfa.	Job Satisfa.	Sleep Duration	Dietary Habits	Degree	Have you ever had suicidal thoughts ?	Work/ Study Hours	Financial Stress	Family History of Mental Illness
2	Male	33	Visakhapatnam	Student	5	0	8.97	2	0	5-6 hours	Healthy	B.Pharm	Yes	3	1	No
8	Female	24	Bangalore	Student	2	0	5.9	5	0	5-6 hours	Moderate	BSc	No	3	2	Yes
26	Male	31	Srinagar	Student	3	0	7.03	5	0	Less than 5 hours	Healthy	BA	No	9	1	Yes
30	Female	28	Varanasi	Student	3	0	5.59	2	0	7-8 hours	Moderate	BCA	Yes	4	5	Yes

MODEL TRAINING AND DETECTION USING SPARK ML

Key Objective:
Build a robust machine learning model capable of predicting depression based on the most relevant features.

- **Dataset Splitting:**

The initial dataset from Kaggle will be divided into two subsets: a training set (majority of the data) and a test set.

- **Model Training:**

We will use Spark ML to train a machine learning model for depression detection based on various features from the dataset.

- **Feature Engineering:**

A correlation heatmap will be created to identify which features have the most significant impact on depression.

- **Testing the Model:**

Once trained, the model will be tested on the test dataset to validate its accuracy in detecting depression among students.

IDENTIFYING KEY INFLUENCING FACTORS ON DEPRESSION

Key Objective:

Determine the most critical factors affecting student depression and categorize them into different groups.

- **Feature Importance Analysis:**

Beyond just predicting if a student has depression, we will analyze which features play a more significant role in determining depression.

- **Categorization of Features:**

The model will classify behavioral parameters (e.g., academic pressure, financial stress) and demographic parameters (e.g., city, living conditions) to evaluate their individual contributions to depression risk.

- **Comparison of Influence:**

This analysis will provide insights into which type of features - behavioral or demographic - have a stronger correlation with depression.

REAL-TIME DEPRESSION DETECTION VIA STREAMING

Key Objective:

Ensure that real-time survey data is structured correctly before streaming via Kafka, allowing Spark ML to process and classify students dynamically.

- **Converting Survey Data Format:**

A new text file containing survey responses from several students will be processed into a CSV or structured text file, aligning with the original dataset's format.

- **Data Streaming with Kafka:**

The newly structured file will be used as input, and Kafka will stream this data in real-time, sending each student's record sequentially.

- **Real-Time Prediction with Spark ML:**

As Kafka streams the data, the trained Spark ML model will process each incoming record in real-time, classifying whether the student has depression.

ACADEMIC ARTICLES

Predicting Student Depression With Measures of General and Academic Anxieties

- How does it related to the subject?

This study explores how various forms of anxiety, including general neuroticism, academic anxiety, and test anxiety, contribute to depression among university students. It provides a psychological framework that helps identify which factors are most predictive of depression, aligning with our goal of analyzing student depression using data-driven techniques.

- What have we learned, and what will we use?

We have learned that academic and test-related anxieties are strong predictors of depression and can serve as early indicators for mental health concerns in students. This insight will guide our feature selection process in Spark ML, ensuring that the most influential variables are emphasized in our model to improve depression prediction accuracy.

ACADEMIC ARTICLES

Prediction of Depression for Undergraduate Students Based on Imbalanced Data by Using Data Mining Techniques

- How does it related to the subject?

This study examines classification techniques and feature selection methods for predicting student depression, focusing on imbalanced datasets. It is relevant to our project as it offers insights into data preprocessing, handling imbalanced data, and optimizing feature selection for accurate predictive modeling.

- What have we learned, and what will we use?

We learned that a technique like SMOTE (Synthetic Minority Over-sampling Technique) can improve model accuracy in depression prediction. Since PySpark lacks native SMOTE support, we'll apply it using Python's "imblearn" library before data loading or use PySpark's oversampling/undersampling methods. Lastly, the paper's feature selection methods - correlation, gain ratio, and Relief - will guide us in identifying key features in Spark ML for better prediction.