PROJECT PROPOSAL
MACHINE LEARNING WITH BIG DATA

PROJECT TITLE

MINING HOSPITAL RECORDS FOR PREDICTING PATIENT DROP-OFF

SUBMITTED BY:

JATIN SHRIVAS [G23AI2094]

SURAJ MOURYA [G23AI2116]

SHUBHAM RAJ [G23AI2028]

SUBMISSION DATE: 31st JAN, 2025

INDIAN INSTITUTE OF TECHNOLOGY, JODHPUR

MINING HOSPITAL RECORDS FOR PREDICTING PATIENT DROP-OFF

Team Composition:

JATIN SHRIVAS [G23AI2094]
SURAJ MOURYA [G23AI2116]
SHUBHAM RAJ [G23AI2028]

APPLICATION OVERVIEW

Hospitals or medical centre are some of the busy place. Victims, Patients ,Sufferer etc come in for appointments, undergo treatments, and (hopefully) follow through on their care plans. But in many cases, they just "drop off" the radar—missing appointments or failing to complete their treatments. That's where our project will steps in. We are trying to build a predictive system using big data and machine learning to identify which victims, patients, sufferer etc are most at risk of dropping off. We will provide real-time, actionable insights to healthcare professionals to catch these red flags early. Using large-scale hospital records like demographics, treatment history, doctor's notes, everything just you name it, we intend to allow hospitals to help increase patient engagement, improve care outcomes, and save resources in the long term so we can help everyone get the right treatment.

OBJECTIVES

Data Collection & Pre-processing

We will gather all relevant hospital records (demographics details, appointment logs, treatment progress, ABHA records, Ayushman Bharat Records, Insurance Details etc) and ensure the data is clean, consistent, and ready for analytics. Missing values, outliers, or messy text data etc will be handle that with robust cleaning pipelines.

Predictive Modelling

Using machine learning models like Logistic Regression, Random Forest, Gradient Boosting, and even Neural Networks, we will try to predict which individual might discontinue treatment. We will compare these models to see which one gives us the most accuracy and best results.

Scalable Deployment

Once our model is ready for prime time, we will make sure it can handle big datasets without collapsing. We will also wrap it into a user-friendly, web-based tool so hospital staff can easily input new data and get predictions instantly.

Real-Time Analytics

In a perfect world, hospitals, medical centres should be able to respond the moment a patient's risk spikes. We want to leverage real time data streaming so our predictive engine is continuously updated, giving healthcare providers immediate insights.

TECH STACKS & JUSTIFICATION

Frontend

- **React.js**: This lets us create a smooth, interactive interface where healthcare providers can see patient risk levels at a glance. Simply for building a responsive and interactive UI.
- **Material-UI**: As a clean, modern look in the frontend is critical to ensure doctors and nurses are not wrestling with outdated designs. Various other Template / Themes/ Packages will be used for rendering dynamic and interactive details and context of the project.

Backend

- **Python (Pandas, Scikit-learn, TensorFlow/PyTorch):** For heavyweight data wrangling and training of the model. For table-like data manipulation, Pandas is used. But whenever it comes to the main ML algorithms, I will try to rely on Scikit-learn. And for deeper leaning TensorFlow or PyTorch whichever possible.
- **Flask:** We will wrap our models in a RESTful API, making it easy to integrate predictions into the frontend or any other service. We can also try to host it on cloud depending on the availability and understanding of the hosting of the AWS.

Big Data Tools

- **Apache Spark**: It will help us distribute the workload across multiple machines, so we are not waiting forever for results when crunching large datasets.
- **Hadoop (HDFS):** Providing a structured way to store and retrieve massive amounts of data.

Cloud Infrastructure

- **AWS (EC2, S3):** Has hosted our application in scalable environments. S3 for data storage, EC2 for the instances in which our application will run.

DELIVERABLES

- **Predictive Model :** A polished ML model (or a set of them) that can label patients as "low risk" or "high risk" for dropping off.
- **Web Application :** Deployed on AWS with an interface accessible to non-technical users. They will be able to login, enter patient information, and see predictions.
- **Codebase and Documentation :** GitHub repo where everything is explained clearly, from setting up the environment to step-by-step deployment.
- **Final Presentation :** We'll put on a live demo, showing exactly how the app works and explaining the real-world impact.

DATA SOURCE, FORMAT, AND SIZE

Data Source: We'll tap into publicly available datasets of Medicare data or public data if possible. If confidentiality constraints apply, we'll create synthetic data that mimics real hospital records.

Data Format:

- **Structured**: CSV files or SQL tables for standard info (name, age, diagnosis codes, appointment logs).
- **Unstructured**: Text notes or doctor's comments for more nuanced insights, requiring some NLP.

Data Size: Anywhere from 2GB to 10GB or more, which is hefty enough to justify using Spark and Hadoop for efficient handling.

KEY TECHNOLOGY CHALLENGES

- **Handling Large-Scale Data :** Hospitals generate loads of data, imagining thousands of patients across multiple years. We'll need distributed computing (Spark) to handle the data gracefully.
- **Ensuring Data Privacy :** Data is sensitive and source of truth. So, we need to respect healthcare regulations and keep our pipelines secure.
- **Model interpretability :** Healthcare providers need to believe in our predictions. If it is too opaque, they might ignore the recommendations. We'll explore feature importance or LIME/SHAP to make it clearer.
- **Real-time predictions :** If we want to catch drop-offs as they happen, our system has to process incoming data and update patient risk scores without missing a beat.

EXPECTED OUTCOME

By the end of this project, we want to have a scalable, efficient solution that warns hospitals or medical institution or medical centres when patients are at risk of discontinuing care. Instead of burdening staff under complicated reports, we will offer a straight forward dashboard and robust predictive models. The hope is that healthcare professionals can intervene earlier, whether that's scheduling a quick follow-up call or adjusting a treatment plan which will improve retention rates and ultimately leading to better patient outcomes.

Thank you