Final Capstone Project Proposal:

Expanding Mental Health Predictive Analysis with Advanced Features and Application

This project builds on the previous mental health capstone by leveraging a new dataset to enhance predictive capabilities - The new dataset is focused on US patients, which offers a more consistent basis for prediction.

The goal is to improve real-world applicability, and accessibility of mental health interventions, incorporating MLOps and ML Engineering practices to ensure robust and scalable solutions.

a. The business problem**:**

   Mental health issues are increasingly prevalent, yet access to timely and personalized support remains a significant challenge. The consequences of inadequate support are not just individual but also have widespread legal and financial impacts.

   Healthcare providers like Kaiser Permanente, the Priory Group, Acadia Healthcare, and Universal Health Services have faced lawsuits, fines, and settlements due to insufficient mental health care, with Universal Health Services paying $122 million for failing to provide proper behavioral health care. These cases underscore the risk and high cost of neglecting mental health support, highlighting the critical need for proactive, comprehensive intervention.

   There is a pressing business need for an effective and scalable solution that predicts and supports mental health needs. By leveraging data-driven insights, early intervention can be ensured, helping to mitigate both personal and organizational risks while fostering better mental health outcomes.

b. Intended stakeholders:

   For **healthcare providers**, the project offers tools to prioritize care and allocate resources effectively. Mental health professionals benefit from enhanced predictive insights to personalize treatments.

   Policymakers can use aggregate data to shape interventions at a community level.

   Finally, **individuals** gain access to mental health resources and support in a more timely manner.

c. Dataset:

The CDC dataset is focused solely on US patients, offering more consistent, representative data for that region. In contrast, the Kaggle dataset was global, but suffered from a high imbalance in the target class across different regions. It was skewed heavily towards the US and a few European countries, leading to challenges in balancing regional data points effectively.

The dataset is primarily sourced from the 2022 and 2023 CDC.gov BRFSS (**Centers for Disease Control and Prevention**, **Behavioral Risk Factor Surveillance System**), offering comprehensive mental health survey data. When combined, these datasets provide an estimated 800,000 data points, offering a robust foundation for building the model.

d. Data science approaches:

Feature Engineering: To transform raw CDC mental health survey data into meaningful inputs for the prediction model by selecting relevant questions on mental health days, lifestyle behaviors, and healthcare access. The goal is to extract key factors that impact mental health outcomes, improving model understanding. This includes combining related questions into single features or categorizing complex responses for better interpretability and predictions.

Predictive Modeling: Using classification algorithms to predict mental health outcomes and identify individuals needing intervention.

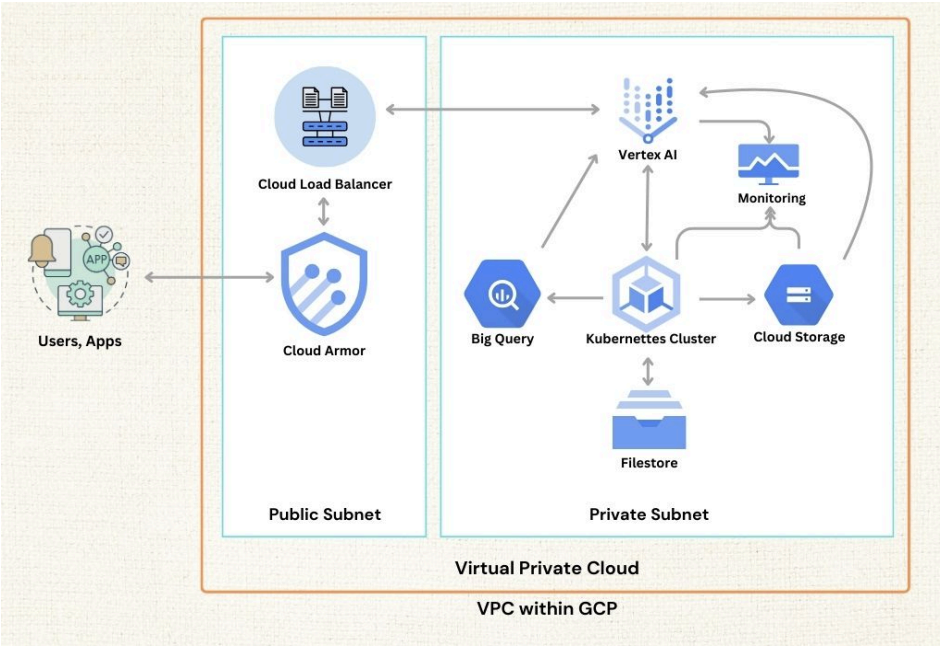
Develop an application that accepts predefined survey inputs based on the most predictive features from the dataset. The app will guide users through a series of questions, collect key inputs, and provide recommendations from the predictive model. The application will be publicly accessible - for testing and evaluation, via a web URL, compatible with any desktop or mobile web browser.

e. MLOps and ML Engineering:

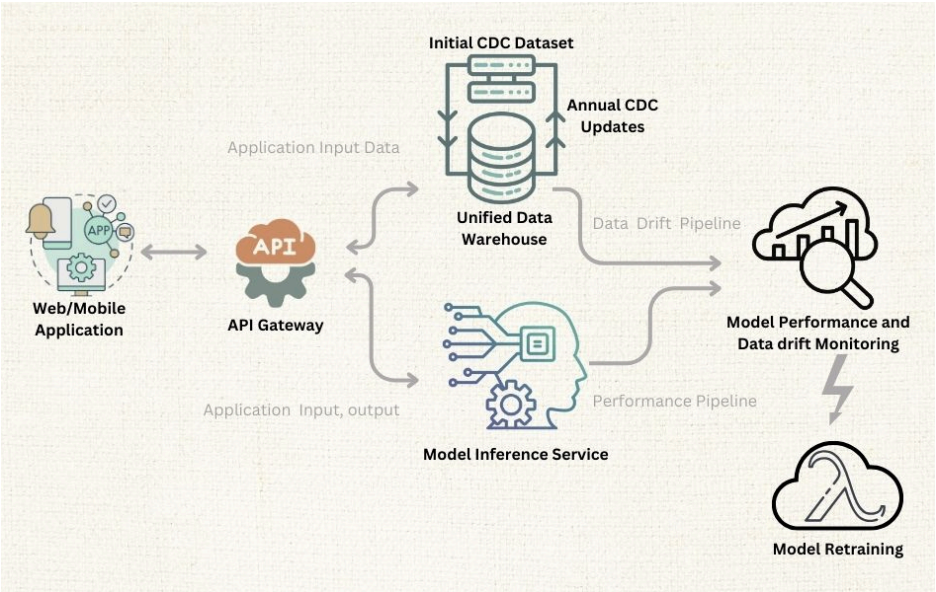
The deployment of the model will follow MLOps best practices to ensure scalability, reliability, and continuous, consistent monitoring, evaluation, updating, and retraining of machine learning models

The application will be deployed as a cloud-based web service on GCP (Google Cloud Platform), accessible via both web and mobile platforms.

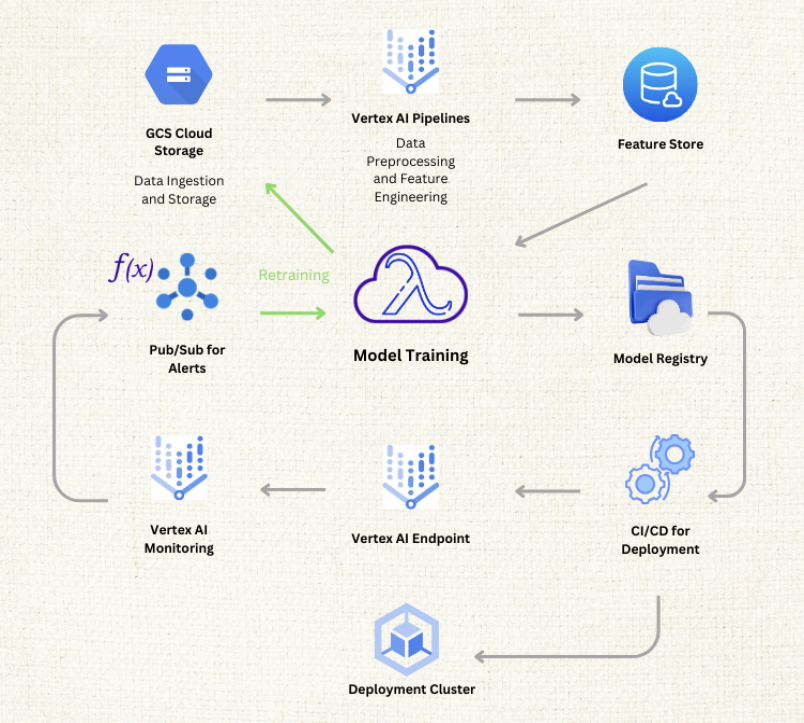
Below are the architecture diagrams for the **Mental Health Support Services**, **Cloud Platform**, and **MLOps** infrastructure.



Google Cloud Architecture Diagram



Mental Health Support Service Architecture Diagram



MLOps Architecture Diagram

Containerization (e.g., Docker) and orchestration tools (e.g., Kubernetes) will be utilized to manage the deployment environment, ensuring smooth and scalable service delivery.

For monitoring, metrics tracking will be implemented, utilizing tools such as Prometheus and Grafana to observe model performance (for instance, F1 Score and Accuracy), data drift, and user behavior (for instance, PSI due to gender and region drift in the new dataset; KL Divergence due to occupation and work interest drift).

Additionally, continuous model development (using CI/CD pipelines) will be applied to retrain and fine-tune the model as new data becomes available, ensuring that the model remains up-to-date and effective over time.

Hyperparameter tuning will also be periodically performed to maintain and potentially improve model performance.

The model fine-tuning and continuous development process will involve the following approaches:

- Retraining when significant data changes occur.
- Retraining when model performance degrades or data drift is detected based on defined metrics.

We will refine and adapt the most suitable approach as part of the ongoing development process.

For simulating data flow over time, we'll use:

Data Replay to simulate realistic scenarios with historical data (See: Enhancing Consistency and Mitigating Bias: A Data Replay Approach for Incremental Learning),

Synthetic Data Generation for varied and stress scenarios (See: Machine Learning for Synthetic Data Generation: A Review), and a

Sliding Window Approach to evaluate models incrementally (See: An adaptive XGBoost-based optimized sliding window for concept drift handling in non-stationary spatiotemporal data streams classifications). These methods ensure a comprehensive evaluation of how the model handles dynamic, real-world data effectively.