Springboard–DSC Program
Capstone Project 3 Proposal

Expanding Mental Health Predictive Analysis with Advanced Features and Application

Mental Health and Well-being: Predicting the Need for Mental Health Support

Prepared by:

Jude M. Santos
Data Scientist

December, 2024

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.

Stakeholders and Relevance

- 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.

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.

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.
We will use logistic regression as our baseline model, along with neural networks, stacked modeling, and other advanced classification algorithms to enhance predictive performance.

Develop an application

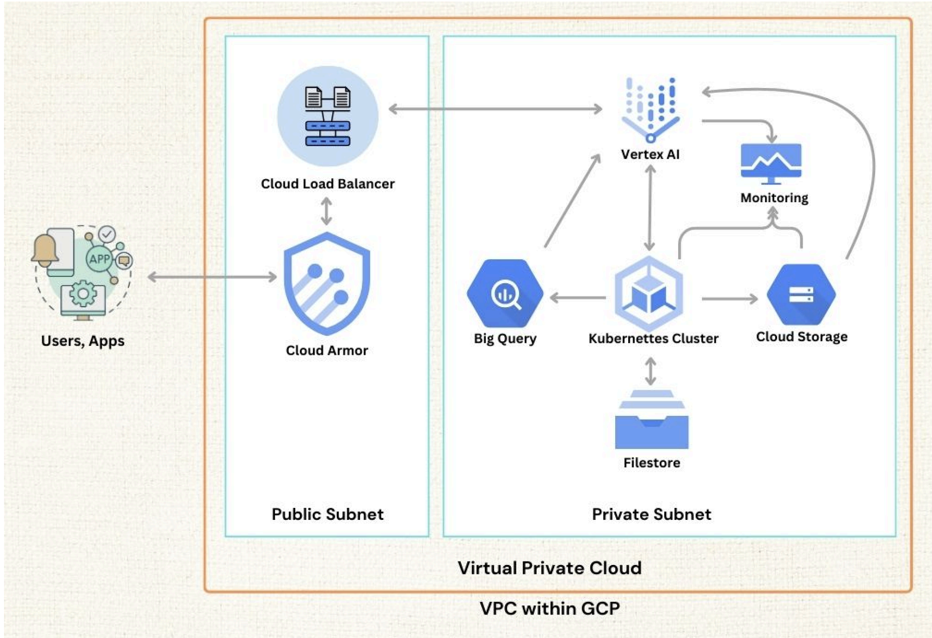
that accepts predefined survey inputs based on the most predictive features from the dataset. The app will guide users through a series of questions, collect key inputs, and provide recommendations from the predictive model. The application will be publicly accessible - for testing and evaluation, via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

MLOps and ML Engineering

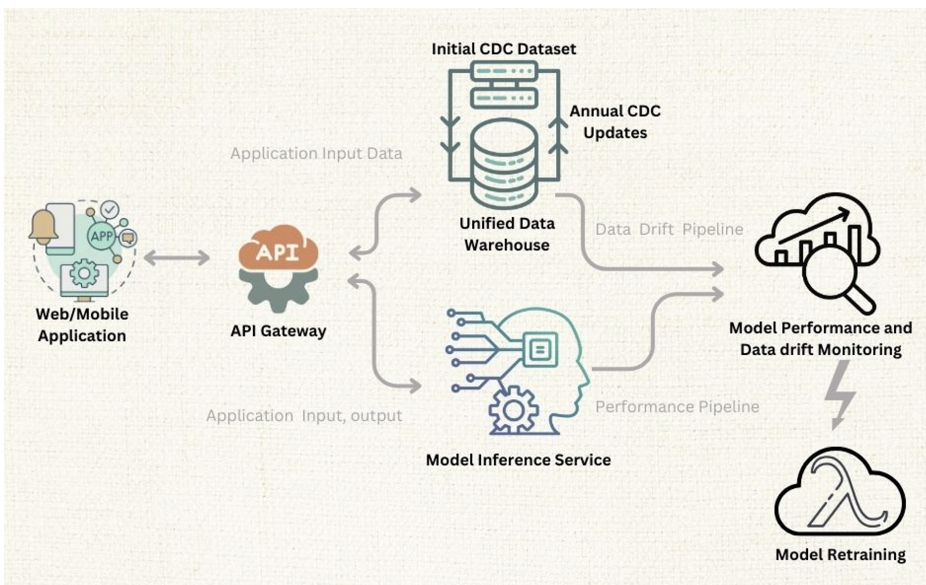
The deployment of the model will follow MLOps best practices to ensure scalability, reliability, and continuous, consistent monitoring, evaluation, updating, and retraining of machine learning models

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

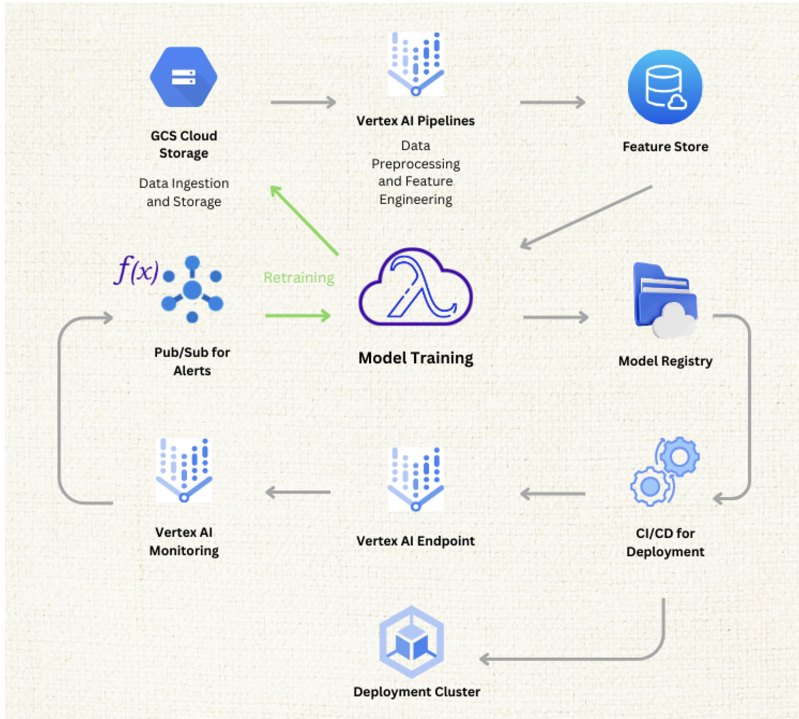
Below are the architecture diagrams for the Mental Health Support Services, Cloud Platform, and MLOps infrastructure.



Google Cloud Architecture Diagram



Mental Health Support Service Architecture Diagram



MLOps Architecture Diagram

Containerization

Using Docker for containerization and GKE Kubernetes as the orchestration tool, the deployment environment will be managed to ensure smooth, scalable, and reliable service delivery.

Monitoring, Metrics

Prometheus and Grafana will be utilized to implement monitoring and metrics, enabling the observation of model performance (e.g., F1 Score and Accuracy), detection of data drift, and analysis of data behavior. This includes tracking PSI (Population Stability Index) for gender and region drift and KL (Kullback-Leibler) Divergence for shifts in occupation and work interest distributions.

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.

Testing and Simulation

- 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: Conditional sliding windows: An approach for handling data limitation in colorectal histopathology image classification). These methods ensure a comprehensive evaluation of how the model handles dynamic, real-world data effectively.

Model Performance Evaluation

We will evaluate the performance of our models using various metrics and validation techniques:

Confusion Matrix and Classification Report

- Provide a comprehensive overview of the model's performance.
- Calculate metrics such as accuracy, precision, recall, and F1-score to assess overall correctness, positive predictive value, sensitivity, and harmonic mean of precision and recall.

Precision-Recall Curves

- Visualize the trade-off between precision and recall.

- Identify the threshold that represents the best compromise between precision and recall.

Cross-validation

- Estimates model hyper-parameters over various partitions of the training dataset.
- Prevents overfitting.
- Uses techniques like k-fold and stratified k-fold.
- Provides reliable performance estimates.

For interpretability

Feature Importance Analysis ranks features with respect to how much their impact in computing predictions.

SHAP (SHapley Additive exPlanations) will provide detailed insights into how individual features contribute to specific predictions, to the model's global performance, and feature interactions.