AAI-501 Group 6

Project Proposal

Title: Multi-Method AI System for Diabetes Risk Prediction and Behavioral Segmentation

1. Problem Statement, Algorithms, and System Overview

This project uses the Diabetes Health Indicators dataset from the UCI Machine Learning Repository to build a system that predicts diabetes risk and identifies health-behavior segments. Diabetes imposes major financial and health burdens, and early identification helps organizations plan effective interventions.

The project applies algorithms:

Supervised Learning: Logistic Regression, Random Forest Classifier, Gradient Boosting Classifier

Unsupervised Learning: K-Means Clustering

Deep Learning: Feedforward Neural Network (MLP)

The system will compare these models on performance, interpretability, and scalability, and discuss how optimization and reinforcement learning concepts could complement decision-making.

2. Related Course Topics

This project aligns with key course areas:

- Foundations of AI and agent design

- State-space and search concepts

- Linear programming and optimization (Simplex/MIP)

- Supervised and unsupervised machine learning

- Deep learning principles

- Full ML lifecycle: preparation, training, evaluation

- Dimensionality reduction (PCA)

- Evaluation metrics: ROC-AUC, precision/recall, F1, Silhouette Score

3. Expected System Behaviors

A. Diabetes Risk Prediction

Outputs the probability of diabetes using health factors such as BMI, activity level, blood pressure, and smoking. Provides interpretability through coefficients, feature importance, or SHAP values.

B. Behavioral/Risk Segmentation

Uses clustering to group individuals with similar health behaviors into segments such as high-BMI sedentary, moderate-risk active, or multi-comorbidity clusters for targeted outreach.

C. Deep Learning Classification

Evaluates whether neural networks capture additional non-linear patterns beyond classical ML models.

4. Key Issues and Analytical Focus

- Class imbalance addressed through class weighting or SMOTE
- Feature correlation handled through PCA or feature selection
- Comparisons of ML, clustering, and deep learning approaches
- Hyperparameter tuning for improved accuracy
- Emphasis on interpretability for healthcare use
- Cluster validation using the Elbow Method and Silhouette Score
- Efficient training on a large dataset (250k+ records)

References

Centers for Disease Control and Prevention. (2016). *Behavioral Risk Factor Surveillance System Survey Data*. U.S. Department of Health and Human Services, CDC.

Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository*. University of California, Irvine. https://archive.ics.uci.edu/

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in Python*. Springer. https://doi.org/10.1007/978-1-0716-2197-4

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica, 31*, 249–268.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Tan, P.-N., Steinbach, M., & Kumar, V. (2019). *Introduction to data mining* (2nd ed.). Pearson.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Zhang, Z. (2016). Variable selection with stepwise and best subset approaches. *Annals of Translational Medicine, 4*(7), 136. https://doi.org/10.21037/atm.2016.03.35