**Predicting Diabetes Risk in the Pima Indian Population Using Machine Learning**

**Sai Shivani Parimi EUID: sp3403**

**Varsha Sane EUID: vs1625**

**Ria Singh EUID: rs1711**

**Description of the Chosen Community/Population:**

The Pima Indians living in Arizona represent a suitable target group for this research because they demonstrate among the highest type 2 diabetes prevalence rates worldwide. A large number of Pima adults suffer from diabetes, according to Knowler et al. (1978) and Looker et al. (2023). Since modern times, the Pima Indians have developed diabetes into a serious public health matter as their native genetics interacted with dietary and lifestyle shifts. The Pima Indians once followed traditional ways of living that included physical activity and food consumption according to their native customs because that was their preferred lifestyle. The introduction of modern developments along with sedentary behaviour patterns, combined with processed foods, resulted in swift increases of metabolic conditions, especially type 2 diabetes (Pickup, 2012).

Insulin resistance, together with impaired glucose metabolism, runs within the genetic structure of Pima Indians according to research findings. The Pima Indians' genetic disposition exposes them to high risk because it interacts with obesity and decreased physical activity to create a wellness disaster, leading to excessive diabetes cases (Pearson, 2015). The Pima Indian population develops diabetes at a younger ages than other populations and shows an average diabetes onset that happens before other populations. The Pima people display both high risk and excellent potential for developing predictive forecasting models that serve early warning and personalized preventive strategies. Our model is designed to help predict diabetes at an early stage for any individuals who are at risk of getting diabetes. So that they can take preventive measures beforehand.

Life course studies from this community enable the collection of essential data required for creating and verifying machine learning algorithms. The long historical data enables scientists to create models that both represent accuracy and conform to cultural practices (Verma & Khatoon, 2024). Choosing the Pima Indian population ensures a strategic opportunity to solve the diabetes epidemic with modern technological approaches.

**Background Information on the Health Problem:**

Type 2 diabetes mellitus (T2DM) appears as a long-term metabolic condition that causes elevated blood glucose because of both insulin resistance and decreased insulin competence. The period of uncontrolled blood sugar creates the risk for significant medical issues including cardiovascular disease, kidney failure, neuropathy and vision loss. The worldwide distribution of T2DM produces substantial healthcare expenses along with causing severe medical conditions that result in numerous deaths.

The Pima Indian population shows increased levels of the problem to an alarming degree. The Pima Indian population experienced type 2 diabetes at rates which were 19 times higher than the incidence rates in comparable white populations according to Knowler et al. (1978). The problem of obesity has continued to worsen through time according to longitudinal studies that identify 50% adult prevalence (Looker et al., 2023).

Numerous elements have led to the high diabetes prevalence such as insulin resistance genetic pre-disposition coupled with diet changes and reduced physical activity towards processed high-calorie foods. The BMI levels of Pima Indians play an essential role in diabetes development since higher body mass index values lead to increased diabetes risk (Pearson, 2015). The prevention of complications requires immediate detection as well as intervention that needs specific targeting in order to decrease healthcare expenses for this population group. Machine Learning models show much potential to identify specific individual risks which enables healthcare providers to start implementing prevention methods at earlier stages of care (M et al., 2024).

Modern machine learning systems can discover hidden medical patterns in data collections that standard statistics systems would miss. Implementation of these technologies becomes essential for effectively fighting diabetes among this particular at-risk population (Stokes et al., 2022).

**Description of the Existing Technology Solutions Available:**

**Blood Glucose Monitoring Kits**:The glucometer represents one type of blood glucose monitoring kit which enables individuals to track their blood sugar levels at home following diagnosis (Grady et al., 2024).

**Continuous Glucose Monitors (CGMs):** The Dexcom G6 and FreeStyle Libre advanced monitoring systems track blood glucose levels around the clock to supply real-time alarms and display data patterns according to (Ashour et al., 2024).

**Diabetes Management Apps**: Patients use mySugr together with Glucose Buddy to document their blood glucose tests as well as nutritional intake combined with physical activities and medication usage (Chang et al., 2022).

**Insulin Pumps**: The system enables patients to track their blood glucose readings with food logs and exercise activities and medication regimen through mySugr and Glucose Buddy (Grady et al., 2024).

**Telehealth and Remote Monitoring**: Patients can track blood glucose measurements together with food records as well as exercise activities and medication intake using Glucose Buddy and mySugr applications (Braune et al., 2020).

**Health Education Programs**: Patients use mySugr and Glucose Buddy to document blood glucose results alongside their eating habits and sport activities and medicine consumption (Mousa et al., 2023).

.

In contrast to traditional reactive approaches, we propose a proactive machine learning (ML) based predictive system to identify individuals at high risk of developing type 2 diabetes, specifically targeting the Pima Indian community.

**Description of the Proposed Technology Solution:**

A proactive machine learning-based predictive system targets the Pima Indian community to identify persons with high type 2 diabetes risk using different methods compared to conventional reactive strategies.

Ensemble learning functions as the proposed technological framework to predict the risk of type 2 diabetes in the Pima Indian population. An ensemble model was built combining predictions from Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Decision Tree, Random Forest, Gradient Boosting, and XGBoost due to the superior predictive outcomes achieved through their merged base model performance. The method achieves better prediction accuracy by combining multiple base models since it decreases the impact of both model variance and model bias.

A predictive system develops robustness through ensemble architecture where multiple algorithms function together either by voting or averaging or boosting. The analysis system evaluates variable importance to identify which factors mainly consist of glucose measurements combined with BMI and age data and insulin data affect the results to the greatest extent.

The platform provides simplified operations for healthcare professionals who receive clear risk assessment information along with detailed contributions regarding risk factor influences. The proposed solution works to deliver early specific treatment interventions which contribute to decreasing diabetes prevalence in the Pima Indian people.

Building an accurate predictive model while maintaining interpretability eliminates the present healthcare analytics challenge for predictive tools which need to perform both functions.

**Explanation of How You Would Solve the Problem with Your Proposed Technology Solution:**

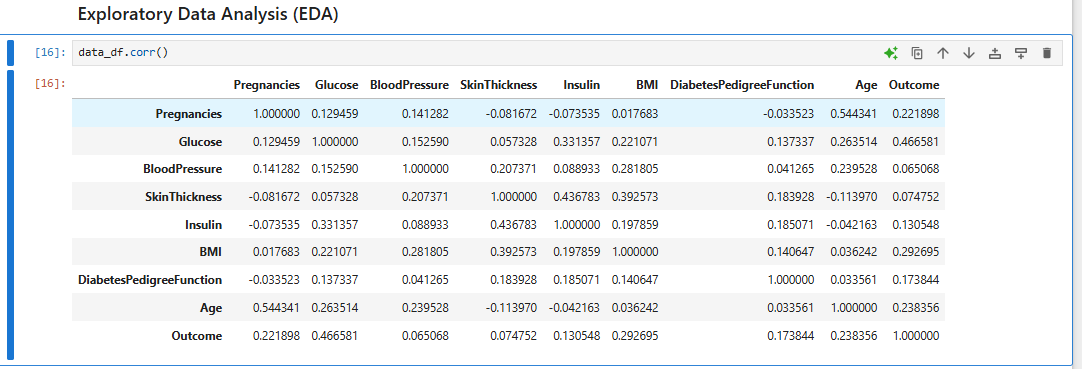
Through its ensemble model technique the system would solve detection deficiencies in the Pima Indian population by conducting personalized risk assessment at an early stage. The system predicts type 2 diabetes risk through personal clinical variables which include glucose levels alongside BMI along with blood pressure as well as insulin levels and family medical information.

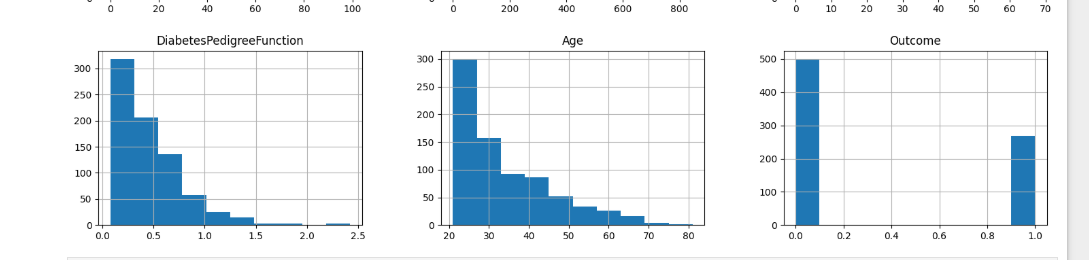
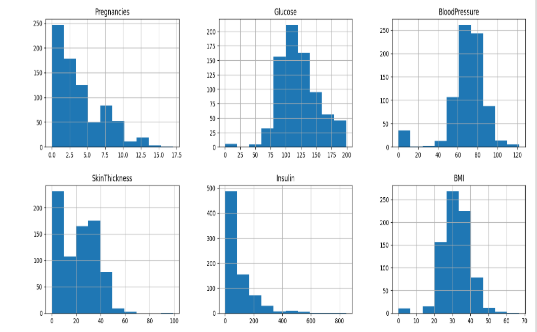
Healthcare providers should use individualized risk scores to find patients who will gain the most from early prevention treatments such as lifestyle education and nutritional advice and exercise plans and medical therapy. Medical prevention measures in preclinical stages may delay or stop the development of diabetes.

The ensemble model demonstrates transparent prediction capabilities through explainable AI techniques which makes its results understandable to healthcare professionals as well as patients. The system develops trust with a clear presentation of its operations which enables joint medical decisions with doctors and their patients. The model can identify high BMI and fasting glucose as the main risk elements for a person which allows healthcare teams to direct interventions toward these priority areas.

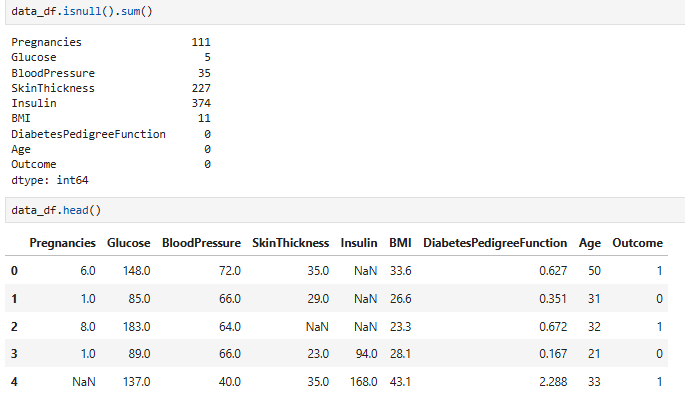
The aggregation of predictions has potential applications in public health through resource planning and identification of high-risk specific populations to develop location-based health programs.

Early detection of diabetes required our proposed machine learning-based predictive system for its solution. Multiple sequential steps made up our proposed solution.

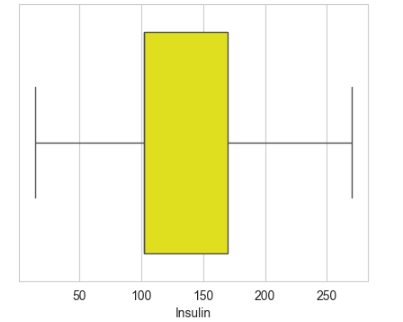
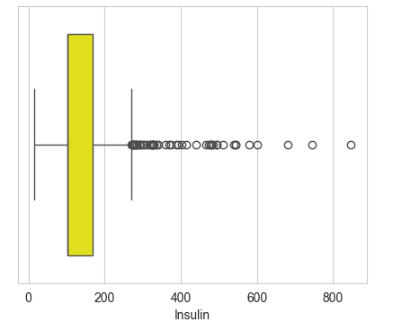
1. **Data Analysis and EDA (Exploratory Data Analysis):** Our team first spent time comprehending all dataset aspects. We analyzed distribution patterns of features together with variable interrelations to detect class imbalance alongside important behavioral trends applicable to diabetes prediction in Pima Indians.  
   

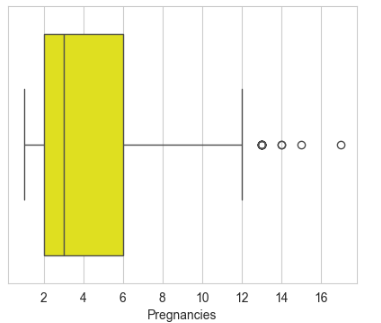
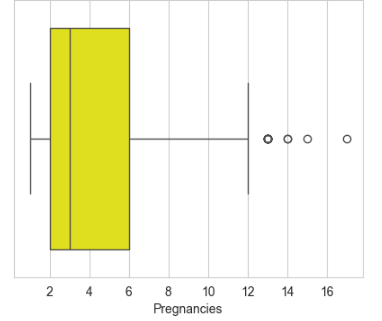


1. **Data Preprocessing:**  
   The missing values received treatment by null imputation approaches in order to complete the dataset before modeling. Evaluation for outliers allowed our team to strip away points that had extreme values which would distort the model learning process or weaken its accuracy.

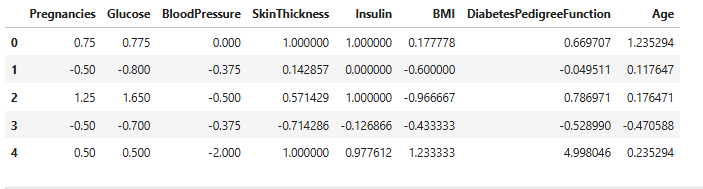


**Outlier Detection**





1. **Feature Engineering:**

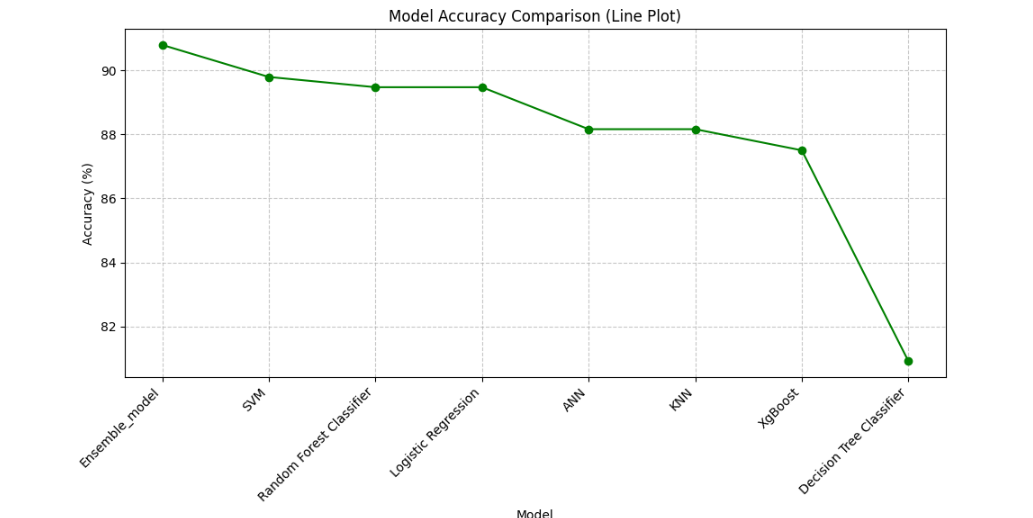
The model received engineered and selected features to improve its diagnostic capability for diabetes cases. The engineering of relevant features during this process enhanced both model interpretability while improving its predictive ability.  


1. **Model Building and Evaluation**:  
   Several machine learning models were trained and evaluated systematically, including:

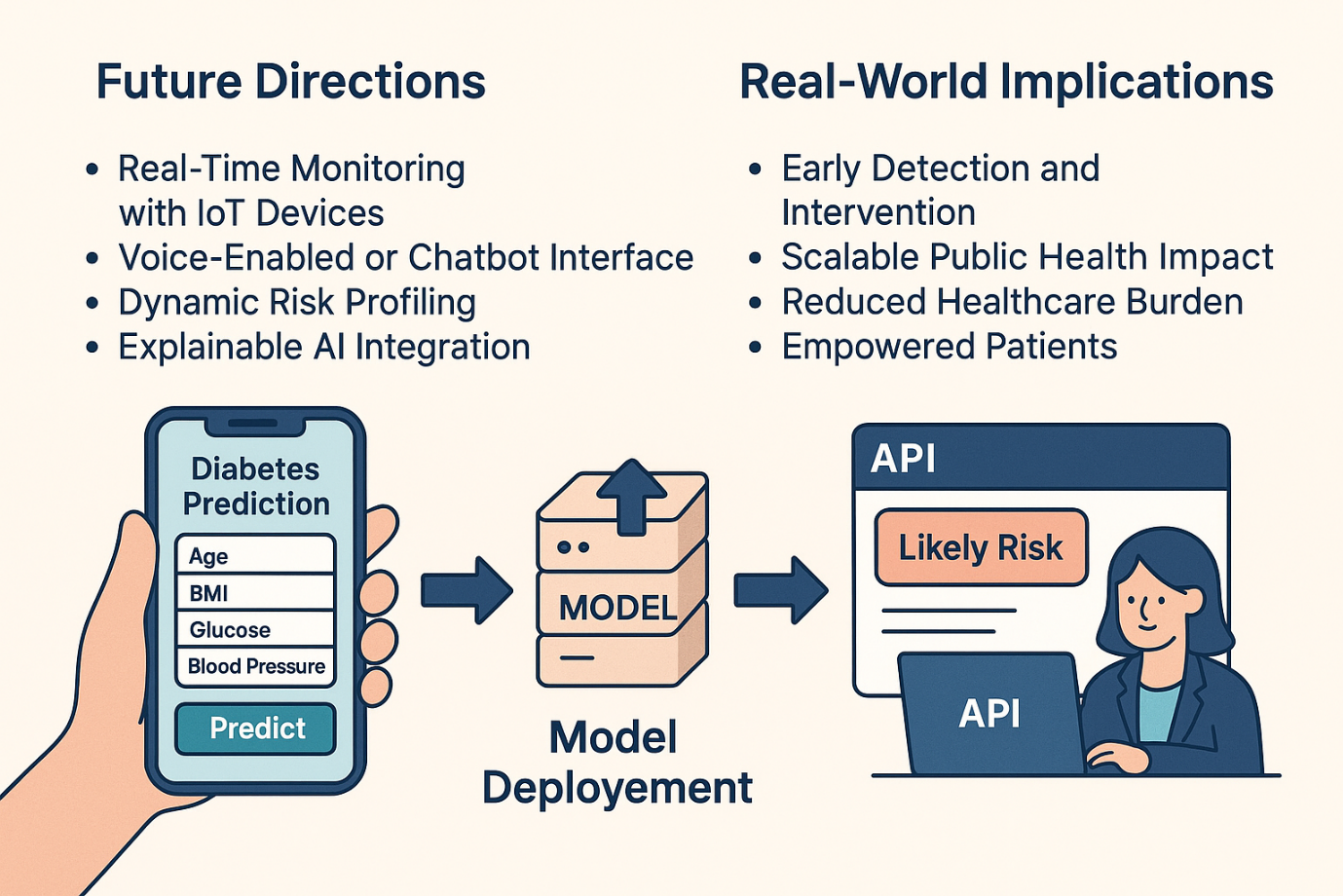
* Support Vector Machines (SVM)
* Random Forest Classifier
* Logistic Regression
* Artificial Neural Networks (ANN)
* K-Nearest Neighbors (KNN)
* XGBoost
* Decision Trees

Each model's performance was assessed based on test accuracy, and results were compared to identify the best-performing approaches.

1. **Final Model Selection (Ensemble Model):**  
   Experimentation produced an ensemble model consisting of several base models giving maximum test accuracy results among all tested candidates.



1. **Deployment Readiness:**  
   The predictive tool is now prepared for deployment because healthcare professionals can use it to identify type 2 diabetes cases at an early stage and initiate intervention for Pima Indians. Healthcare professionals can input patient health information into the app or web interface to receive real-time predictions indicating the risk of developing diabetes. Based on these predictions, they can decide on early interventions to help reduce or manage that risk.



**Contributions:**

**Shivani**: Data preprocessing, exploratory analysis, and feature engineering.

**Varsha**: Machine learning implementation, model evaluation, and hyperparameter tuning.

**Ria**: Explainable AI integration, user interface development, and documentation.

**References**

Ashour, A. F., Fouda, M. M., Fadlullah, Z. M., & Ibrahem, M. I. (2024). Optimized Neural Networks for Diabetes Classification Using Pima Indians Diabetes Database. IEEE, 1–7. <https://doi.org/10.1109/icmi60790.2024.10585703>

Braune, K., Boss, K., Schmidt-Herzel, J., Gajewska, K. A., Thieffry, A., Schulze, L., Posern, B., & Raile, K. (2020). Shaping workflows in digital and remote diabetes care during the COVID-19 pandemic via service Design: Prospective, longitudinal, open-label feasibility trial. JMIR Mhealth and Uhealth, 9(4), e24374. https://doi.org/10.2196/24374

Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications, 35(22), 16157–16173. <https://doi.org/10.1007/s00521-022-07049-z>

Grady, M., Cameron, H., & Holt, E. (2024). Improved Glycemic Control Using a Bluetooth®-Connected Blood Glucose Meter and a Mobile Diabetes App: Real-World Evidence From Over 144 000 People With Diabetes. Journal of Diabetes Science and Technology. https://doi.org/10.1177/19322968221148764

Knowler, W. C., Bennett, P. H., Hamman, R. F., & Miller, M. (1978). DIABETES INCIDENCE AND PREVALENCE IN PIMA INDIANS: a 19-FOLD GREATER INCIDENCE THAN IN ROCHESTER, MINNESOTA. American Journal of Epidemiology, 108(6), 497–505. <https://doi.org/10.1093/oxfordjournals.aje.a112648>

Looker, H. C., Chang, D. C., Baier, L. J., Hanson, R. L., & Nelson, R. G. (2023). Diagnostic Criteria and Etiopathogenesis of Type 2 Diabetes and Its Complications: Lessons from the Pima Indians. Presse Medicale (Paris, France : 1983), 52(1), 104176. <https://doi.org/10.1016/j.lpm.2023.104176>

M, M. B., A, H., BT, I., & S, N. A. (2024). Exploratory Data Analysis and Predictive Modeling of Pima Indian Diabetes. IEEE, 1–6. <https://doi.org/10.1109/iacis61494.2024.10721874>

Mousa, A., Mustafa, W., Marqas, R. B., & Mohammed, S. H. M. (2023). A comparative study of diabetes detection using the PIMA Indian Diabetes Database. The Journal of the University of Duhok, 26(2), 277–288. <https://doi.org/10.26682/sjuod.2023.26.2.24>

Pearson, E. R. (2015). Dissecting the etiology of Type 2 diabetes in the PIMA Indian population. Diabetes, 64(12), 3993–3995. <https://doi.org/10.2337/dbi15-0016>

Pickup J. C. (2012). Insulin-pump therapy for type 1 diabetes mellitus. The New England journal of medicine, 366(17), 1616–1624. https://doi.org/10.1056/NEJMct1113948

Stokes, T., Wilkinson, A., Jayakaran, P., Higgs, C., Keen, D., Mani, R., Sullivan, T., Gray, A. R., Doolan-Noble, F., Mann, J., & Hale, L. (2022). Implementation of the Diabetes Community Exercise and Education Programme (DCEP) for the management of type 2 diabetes: qualitative process evaluation. BMJ Open, 12(5), e059853. https://doi.org/10.1136/bmjopen-2021-059853

Verma, P., & Khatoon, A. (2024). Data Mining Applications in Healthcare: A Comparative Analysis of Classification Techniques for Diabetes Diagnosis Using the PIMA Indian Diabetes Dataset. IEEE, 1–5. <https://doi.org/10.1109/iciptm59628.2024.10563296>