AI-DRIVEN APPLICATION FOR DIABETES CARE: PREDICTIVE ANALYSIS AND PERSONALIZED RECOMMENDATIONS

Damini Prashant Vichare, Manisha Lagisetty, Sai Charishma Kurmala, Yamini Muthyala, Yuting Sha

Motivation:

The escalating global incidence of diabetes underscores the critical need for enhanced early detection methods that can be accessed independently by the public. The traditional approach to diagnosing diabetes, which typically involves clinical assessments and laboratory tests, presents substantial challenges in terms of expense, physical accessibility, and time constraints. These barriers can lead to delayed diagnosis and potentially exacerbate health outcomes. Moreover, the absence of immediate and reliable health advice prevents many individuals from effectively managing their risk factors and making informed decisions about their health. This project is motivated by the necessity to bridge these gaps through a straightforward, user-centric solution that empowers individuals to assess their diabetes risk and receive actionable health guidance promptly. By providing a tool that simplifies the diabetes screening process and offers immediate dietary and lifestyle advice, this initiative aims to facilitate proactive health management and improve the overall effectiveness of diabetes prevention strategies, reducing the burden of this pervasive disease.

Background:

Diabetes has emerged as a formidable global health crisis, affecting approximately 537 million adults worldwide, a figure that represents about one in ten individuals aged 20-79 years. This number is projected to escalate to 643 million by 2030 and 783 million by 2045. The prevalence is disproportionately higher in low and middle income countries, where over three-quarters of diabetic adults reside. In 2021 alone, diabetes was responsible for 6.7 million deaths, roughly translating to one death every five seconds, and incurred a staggering health expenditure of at least USD 966 billion, a 316% increase in the last 15 years. The burgeoning impact of this chronic disease highlights an urgent need for innovative solutions to mitigate its global burden through more accessible and effective early detection and management strategies.

Literature review:

The urgent need for advanced predictive models for early diabetes detection has been the focus of several recent studies, each exploring various data mining techniques to address this pervasive health issue. For instance, a retrospective cohort study by Zhang et al. (2022) demonstrated the utility of data mining in early diabetes prediction by analyzing electronic health records (EHRs) to identify at-risk individuals. Similarly, research by Ali et al. (2018) employed multiple machine learning algorithms to predict diabetes, showing promising accuracy levels. Meanwhile, Sharma et al. (2017) focused on the application of several data mining methods to predict diabetes, emphasizing the effectiveness of decision trees and ensemble methods.

Despite these advancements, many existing models suffer from limitations such as reliance on incomplete datasets with efficient data mining techniques and inadequate focus on user-friendly interfaces that could aid individuals in understanding their health status proactively. Our project stands out in the realm of data mining for diabetes prediction by integrating advanced sampling techniques and optimization of classification algorithms to address the critical challenge of class imbalance - a prevalent issue often overlooked in standard predictive models. Our approach enhances model sensitivity and precision through targeted data preprocessing and the strategic use of SMOTE, ADASYN, and RandomOverSampler with ensemble and classification classifiers. This method not only improves detection rates of diabetic cases, as evidenced by our enhanced recall and F1 scores but also reduces the risk of false positives, a crucial aspect in medical diagnostics. This comprehensive and balanced approach ensures our model's superiority in both predicting diabetes accurately and ensuring the model's applicability in real-world clinical settings.

Our application also incorporates a real-time user-friendly interface allowing for immediate, personalized risk assessments in predicting diabetes and health advice that provides tailored dietary and lifestyle recommendations, enhancing user engagement and promoting proactive health management.

Methodology:

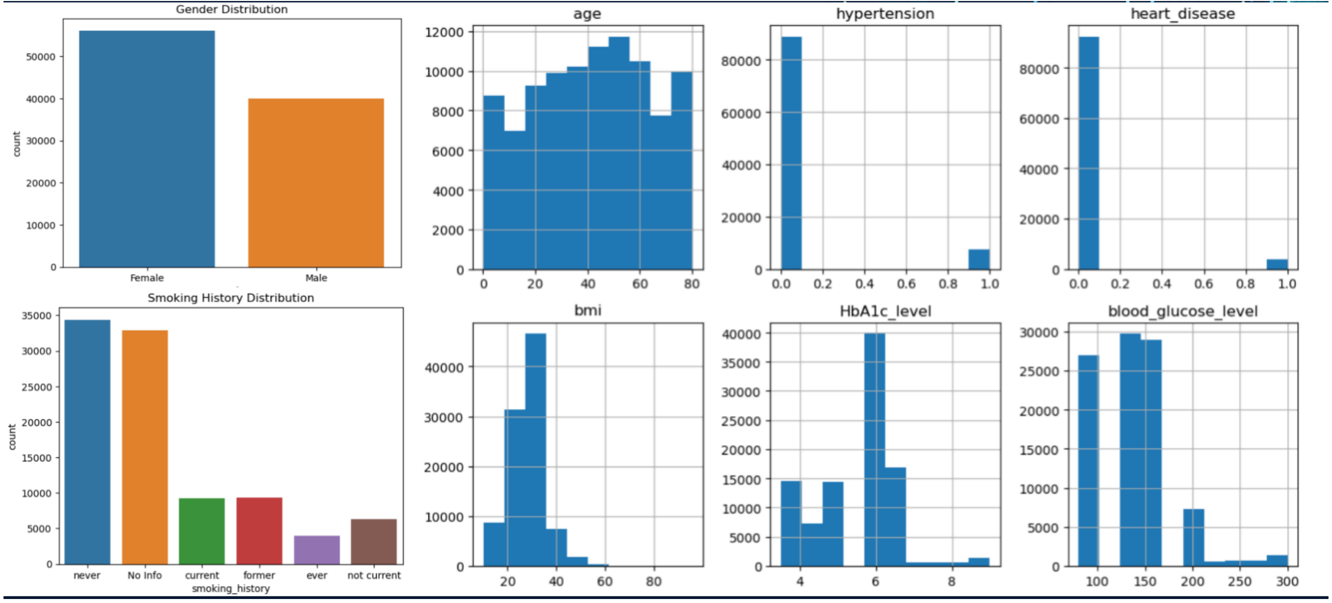
Data Collection Process:

The diabetes dataset is obtained via Diabetes prediction dataset, which consists of 100,000 records and eight features, including gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, and blood glucose level. The target variable "diabetes" is binary, with 1 indicating diabetes and 0 indicating no diabetes, and the dataset is imbalanced, with 91.2% non-diabetic and 8.8% diabetic cases.

Exploratory Data Analysis:

In this part, we aim to understand how variables are distributed, identify any outliers or anomalies, and explore relationships between different features. Firstly, we checked the quality of the data, which involved checking for null and missing values and removing duplicate data points. Then we performed several data visualizations to understand our data patterns and gain insights from them. Figure 1 illustrates the univariate analysis, displaying the data distribution of each feature.

Figure 1: Univariate Analysis



We also explored the underlying relationship between the feature and the target variable. Figure 2.1 displays the distribution of diabetes by gender. Figure 2.2 presents the proportion of diabetes and non-diabetes across different smoking histories, revealing that individuals who have never smoked exhibit the highest proportion of diabetes. Figure 2.3 shows the distribution of diabetes across different age groups, indicating that the age group 60-80 has a higher proportion of diabetes. In Figures 2.4 and 2.5, the density of diabetes across different blood glucose and HbA1c levels is depicted, respectively. Figure 2.6 shows the BMI distribution between the diabetes and non-diabetes groups, highlighting that the diabetes group tends to have higher BMI values.

Figure 2.1 Gender Vs Diabetes

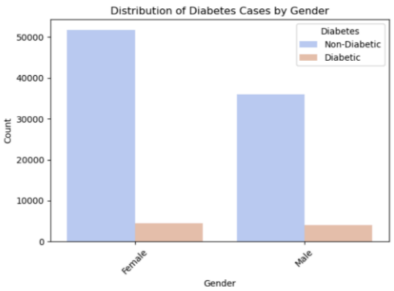


Figure 2.2 Smoking History Vs Diabetes

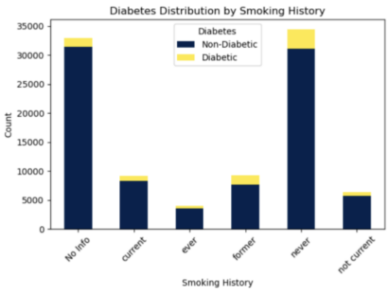


Figure 2.3 Age Vs Diabetes

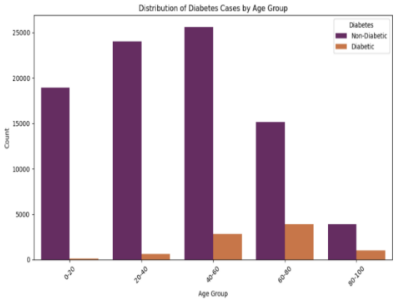


Figure 2.4
Blood Glucose Vs Diabetes

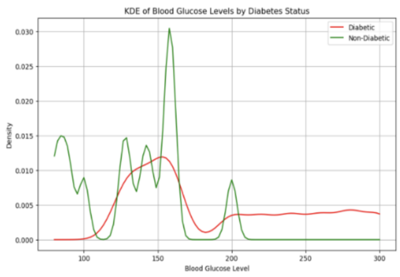


Figure 2.5
HbA1c Level Vs Diabetes

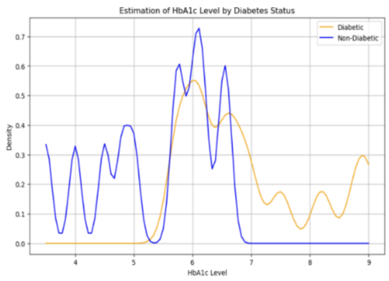
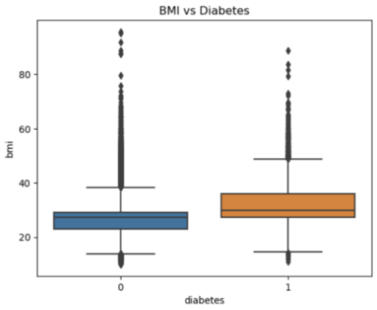


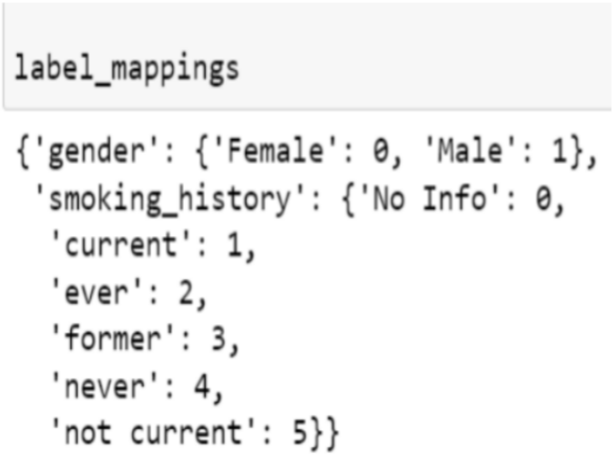
Figure 2.6 BMI Vs Diabetes



Data Pre-Processing:

In the data preprocessing step, we transform the data to make it suitable for further modeling and analysis. Among the 8 features, two of them are categorical features: gender and smoking history. To incorporate these categorical features into our analysis, we convert them into numerical features using label encoding. Additionally, we save the label encoding mapping to ensure interpretability after modeling. Figure 3 shows the details of label encoding mapping. Then, we plotted the correlation matrix, depicted in Figure 4, ensuring that we will not use correlated features in our analysis.

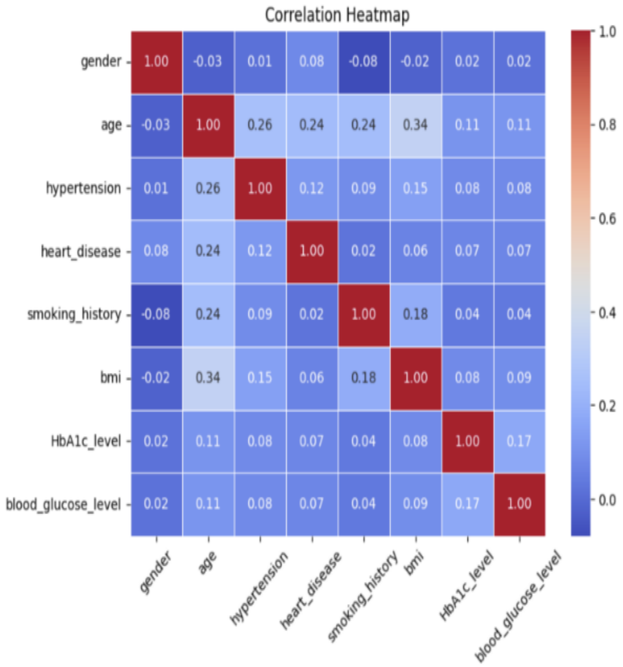
Figure 3
Label Encoding Mapping



```
label_mappings
```

```
{'gender': {'Female': 0, 'Male': 1},
 'smoking_history': {'No Info': 0,
  'current': 1,
  'ever': 2,
  'former': 3,
  'never': 4,
  'not current': 5}}
```

Figure 4
Correlation Matrix



	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level
gender	1.00	-0.03	0.01	0.08	-0.08	-0.02	0.02	0.02
age	-0.03	1.00	0.26	0.24	0.24	0.34	0.11	0.11
hypertension	0.01	0.26	1.00	0.12	0.09	0.15	0.08	0.08
heart_disease	0.08	0.24	0.12	1.00	0.02	0.06	0.07	0.07
smoking_history	-0.08	0.24	0.09	0.02	1.00	0.18	0.04	0.04
bmi	-0.02	0.34	0.15	0.06	0.18	1.00	0.08	0.09
HbA1c_level	0.02	0.11	0.08	0.07	0.04	0.08	1.00	0.17
blood_glucose_level	0.02	0.11	0.08	0.07	0.04	0.09	0.17	1.00

After that, we divided the dataset into training, validation, and testing sets at a ratio of 70:20:10. The training, validation and test set contains 67,289 samples, 19,226 samples, and 9,613 samples, respectively. Until now, the data is well prepared for modeling.

Modeling, Model Details and Training:

We used K-means clustering, Naïve Bayes, Random Forest and XGBoost in our modelling process. The first method we chose is K-means clustering, despite our target problem being a traditional classification problem. This choice is motivated by our future plan to integrate the model into a real-time application, where response time is crucial. This is the primary reason we choose clustering to address the diabetes prediction problem. Since our dataset contains ground truth labels, we would utilize supervised similarity-oriented evaluation measures such as the Rand statistic and Jaccard correlation to validate the clustering performance.

The machine learning algorithms we applied were determined by popular and higher-performing models highlighted in the literature survey. We conducted the modeling on both standardized and unstandardized data. Given the significant class imbalance in our dataset, which may affect model performance, we utilized up sampling techniques such as SMOTE, ADASYN, and Random Oversampling to address this imbalance issue. Additionally, various evaluation metrics such as F1 score, recall, sensitivity, and accuracy were employed. Given the medical nature of the problem, we put more attention on the recall score. During the EDA stage, we observed distinct characteristics among different groups, such as different genders. We will stratify the data based on the confounding variable to factor out the effect of the confounders.

Experiments and Results:

We conducted rigorous experiments to optimize our diabetes prediction tool, focusing on maximizing recall to enhance its sensitivity in identifying diabetes. We prioritized recall as it ensures that our model correctly identifies as many actual cases of diabetes as possible, minimizing the risk of missing patients who require intervention.

These experiments and evaluations were conducted during various approaches listed below:

- **Baseline Models:** Initially, models were trained on the unbalanced dataset to establish a baseline for performance comparison.
- **Standardization:** The data was then standardized to normalize the range of continuous input variables, which helps improve the performance of our models.
- **Balancing Techniques:** To address the issue of class imbalance, over-sampling techniques like SMOTE, ADASYN and RandomOverSampler were applied post-standardization. These methods were used to generate synthetic samples for the minority class, aiming to balance the dataset and improve model fairness.

Evaluation Metrics Used:

In medical diagnostics, accurately identifying diseases like diabetes is crucial. Here’s a concise explanation of key metrics that we used to evaluate our diabetes prediction model:

- **Recall (Sensitivity):** Ensures no diabetic case is missed, crucial for preventing complications from undiagnosed diabetes.
- **Precision:** Important for minimizing false positives, thereby avoiding undue stress and unnecessary medical procedures.
- **Accuracy:** Measures the model's effectiveness in identifying diabetic and non-diabetic cases, though it can be misleading in imbalanced datasets.
- **F1-Score:** Combines precision and recall into a single metric, critical in imbalanced scenarios to ensure fair representation of both classes.
- **ROC-AUC:** Evaluates the model’s ability to differentiate between classes at various thresholds, with a higher AUC indicating better performance.

Baseline Model Performance (Without Standardization and Sampling):

The primary objective of this phase was to evaluate how well our proposed models perform without any data preprocessing interventions such as standardization or balancing techniques.

Techniques	Models	Recall (Sensitivity)	Precision	Accuracy	F1-Score	ROC AUC
Unbalanced (Without Standardization)	Random Forest	**0.69**	0.92	0.97	0.79	0.96
	XGBoost	**0.69**	0.94	0.97	0.80	0.97
	Naive Bayes	**0.64**	0.52	0.92	0.57	0.93

The baseline performance of our models without standardization or sampling indicates moderately high recall and accuracy, particularly for Random Forest and XGBoost, both achieved 0.69 recall and 0.97 accuracy. These models also exhibit strong precision and F1-scores, with XGBoost slightly outperforming Random Forest in precision and F1-score. Naive Bayes, while less accurate at 0.92, shows lower precision and F1-score, highlighting its struggles with the imbalanced dataset. Standardization is essential next to ensure uniform data scales, enhancing model performance and fairness, especially in handling features differently affected by class imbalances.

Model Performance with Data Standardization:

In this phase, we used a StandardScaler to normalize feature values across the dataset, reapplying our models to evaluate the impact of feature scaling on learning efficacy and generalization to unseen data.

Technique	Models	Recall	Precision	Accuracy	F1-Score	ROC AUC
Standardized Data	Random Forest	**0.69**	0.92	0.97	0.79	0.96
	XGBoost	**0.69**	0.94	0.97	0.80	0.97

The standardized data application reveals consistent patterns from the baseline for Random Forest and XGBoost, showing only modest improvements in recall and F1 scores. The Random Forest demonstrates an ongoing struggle with recall at 68.52%, while XGBoost shows slight improvements, highlighting challenges with the precision-recall trade-off in the face of class imbalances. The Naive Bayes model was not standardized due to its assumption of normally distributed data. Standardization has enhanced the learning effectiveness of Random Forest and XGBoost, helping normalize feature scales, yet it did not fully address the class imbalance, crucial for improving minority class prediction.

Evaluating Sampling Techniques on Standardized Data:

Due to dataset imbalance biasing models towards the majority class, we applied sampling techniques post-standardization to improve predictive performance for the minority (diabetic) class.

Technique	Models	Sampling Method	Recall	Precision	Accuracy	F1-Score	ROC AUC
Standardized Data with Sampling Techniques	Random Forest	SMOTE	**0.74**	0.77	0.95	0.75	0.96
		ADASYN	**0.77**	0.72	0.95	0.75	0.96
		RandomOverSampler	**0.71**	0.83	0.96	0.77	0.96
	XGBoost	SMOTE	**0.72**	0.91	0.96	0.80	0.97
		ADASYN	**0.73**	0.88	0.96	0.80	0.97
		RandomOverSampler	**0.87**	0.52	0.91	0.65	0.97
	Naive Bayes	SMOTE	**0.80**	0.41	0.87	0.54	0.88
		ADASYN	**0.89**	0.32	0.82	0.47	0.84
		RandomOverSampler	**0.78**	0.41	0.88	0.54	0.89

The application of sampling techniques after standardizing data markedly improved recall from 0.64 to 0.87 specifically in XGBoost which is crucial for detecting the minority diabetic class. Random Forest and XGBoost showed significant F1 score improvements, up to 0.77 and 0.80 respectively, enhancing their ability to identify diabetic cases. While Naive Bayes achieved high recall, its precision dropped, underscoring the trade-offs in managing class imbalance. This strategy effectively minimized majority class bias, boosting model accuracy in diagnosing diabetes.

Saved our best model:

After carefully evaluating, we selected the XGBoost model with RandomOverSampler applied to standardized data as our optimal model for diabetes prediction. This decision was driven by the model's impressive ability to balance sensitivity. Specifically, the XGBoost model demonstrated a high recall of 87% for the diabetic class, indicating superior detection of positive cases, crucial for medical diagnostics. Our model effectively addresses the critical need for high recall in healthcare applications, prioritizing the detection of all potential diabetic cases.

Influence of Confounding variable:

We identified 'age' as a confounding variable in our diabetes prediction model due to its strong correlations with key features like BMI and hypertension as observed in the correlation heatmap and its strong association with the diabetes outcome. To address this, we stratified our dataset by age, categorizing it into specific groups to ensure equitable model performance evaluation across different age demographics. This stratification aimed to provide proportional representation in the training, validation, and test sets, enhancing prediction accuracy and reducing bias. Despite these efforts, training an XGBoost model with Random Over Sampler on the stratified data revealed no significant improvement in recall compared to non-stratified data, suggesting that the model already adequately captures diabetes risk prediction.

Application Development:

Our application, the Diabetes Prediction Center, offers an intuitive interface that allows users to input personal health data to receive instant diabetes risk assessments. Integrated with our best-performing predictive model, the platform also features a lifestyle application for real-time health-related advice and a BMI calculator to measure body mass index. This setup enhances user engagement and supports preventive health measures effectively.

Discussion and Future Improvements:

Our Diabetes Prediction Center, a robust tool for predicting diabetes risk, is poised for significant enhancements to broaden its capabilities. Currently, the tool effectively leverages sophisticated machine learning models to assess diabetes risk from user-provided health data. The next phase in our development process involves refining these models to differentiate more accurately between type 1 and type 2 diabetes. This distinction is crucial, as each diabetes type has different causes, risk factors, and management strategies. To achieve this, we plan to incorporate advanced algorithms and expand our datasets to include not only traditional health indicators but also genetic markers, autoimmune antibodies, and in-depth lifestyle data. By using this data, we are planning to leverage more sophisticated data mining techniques such as power analysis, clustering to identify and analyze subgroups within the diabetes spectrum that exhibit similar characteristics or respond similarly to treatments. This approach aims to further tailor our diagnostics and interventions, enhancing the tool's clinical relevance.

Looking ahead, we are set to integrate our application with the latest IoT devices, including smartphone apps and innovative diabetic body patches that monitor glucose levels in real time. This integration promises to revolutionize how data flows into our models, significantly enhancing the accuracy of our predictions. Real-time data acquisition allows for immediate adjustments to lifestyle or medical interventions, providing users with timely feedback that could prevent severe diabetic episodes or complications. Furthermore, we plan to employ association rule mining to explore and identify relationships between various lifestyle factors and their influence on diabetes risk. This technique will help reveal hidden patterns that will inform the development of targeted interventions and customized educational resources. We will be enhancing our application to include features for direct communication and collaboration between users and healthcare providers, such as appointment scheduling, medication management, and secure messaging. These improvements aim to make healthcare management more efficient and user-friendly. By fostering better interaction and a proactive health community, and integrating advanced data mining techniques, we aim to significantly improve diabetes management and preventive health strategies, tailoring our tool to meet specific user needs.