

Group Project
Diabetes Prediction – Final

Introduction to Artificial Intelligence
(BDM 3014)

Yorbis Daniel Alarcon (C0941168)

Hazel Santons (C0915982)

Komal Nandal (C0933116)

Sangsun Lee (C0905412)

Gustavo Vera Suarez (C0917164)

Sneha Painli (C0933116)

Mahek Ghanchi (C0937254)

Krunal Patel (C0936008)

Model interpretability and tuning	3
1. Importance of Blood Pressure Imputation in Diabetes Prediction	3
Using LIME for Imputation Model Interpretability	3
2. Feature Selection for Imputation Models	3
3. Evaluating Feature Sets for Imputation	4
4. Adding New Features (with Reference to Research like PIMA Indian Dataset)	4
Why Were New Features Added?	4
Incorporating Expanded Age Range	4
References to PIMA Indian Dataset	5
Key Insights from PIMA Research:	5
Impact of New Features	5
Concluding Implications	5
5. Lime Results for the Updated Features	6
6. Division of Feature Sets	6
7. Model Tuning Results	6
8. Hyperparameter tuning	7
9. Insights from Feature Expansion and Model Tuning	8
Balancing the Dataset with Male Feature Inclusion	9
Model System Development	11
Overview of the Dual-Model System	11
Primary Model: Diabetes Prediction	11
Secondary Model: Risk Stratification	11
Synergy Between the Models	11
Insights from SHAP and LIME	12
Conclusion	12
MECE Table	13
1. Problem Definition	13
2. Data Collection and Source	13
3. Data Preprocessing	14
4. Feature Selection and Evaluation	14
5. Model Development	15
6. Model Performance	15
Project Work Table	17
Project Board	19
Git Repository	20

Model interpretability and tuning

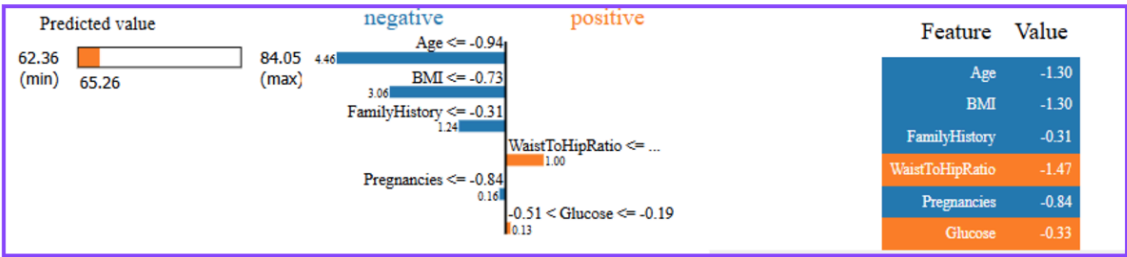
1. Importance of Blood Pressure Imputation in Diabetes Prediction

Blood pressure is a key indicator of metabolic health and a significant factor in diabetes prediction models. Since our dataset had a substantial proportion of missing blood pressure values, it was critical to impute these values accurately to avoid introducing noise or bias into the subsequent diabetes prediction model.

Using LIME for Imputation Model Interpretability

LIME (Local Interpretable Model-Agnostic Explanations) was applied to the imputation model to identify the most influential features contributing to the predicted blood pressure values. This interpretability step ensured that the imputation model was not only statistically effective but also aligned with clinical expectations regarding the relationships between features.

2. Feature Selection for Imputation Models



Based on insights from the **LIME plot**, which highlighted the positive and negative contributions of various features, three feature sets were defined to evaluate the optimal combination for imputing missing blood pressure values. **Set 1** included ['Age', 'BMI', 'WaistToHipRatio'], focusing on the key predictors identified as significant contributors to blood pressure variability. **Set 2**, consisting of ['Age'], was designed to explore the individual impact of age on the imputation process. **Set 3**, a comprehensive set, incorporated all available predictors except the target, including ['Age', 'BMI', 'WaistToHipRatio', 'Pregnancies', 'Glucose', 'FamilyHistory'], to provide a holistic view of blood pressure determinants. The LIME analysis revealed that **WaistToHipRatio** had the strongest positive influence, underscoring its critical role in explaining blood pressure variability. Meanwhile, **Age** and **BMI** were identified as strong negative contributors, indicating their inverse relationship with blood pressure and their necessity in feature selection. Although **Pregnancies**, **Glucose**, and **FamilyHistory** showed smaller contributions, their complementary value supported their inclusion in the comprehensive feature set. These insights informed the systematic evaluation of feature sets and optimized the imputation process.

3. Evaluating Feature Sets for Imputation

Missing values successfully filled!

Chosen Feature Set for Imputation of BloodPressure:
['Age', 'BMI', 'WaistToHipRatio']

Feature Set	R-squared (Validation)	MSE (Validation)	R-squared (Test)	MSE (Test)
Set 1 (Age,BMI,WaistToHipRatio)	0.094433	100.519790	0.086290	102.787343
Set 2 (Age Only)	0.073139	102.883468	0.066154	105.052557
Set 3 (All Features Except Outcome)	0.094287	100.535924	0.088351	102.555416

To evaluate the effectiveness of each feature set, imputation models were developed and tested using the defined sets, with performance measured through R-squared and Mean Squared Error (MSE). **Set 1**, which combined **Age**, **BMI**, and **WaistToHipRatio**, achieved slightly better performance with an R-squared value of **0.086290** on the test set, indicating that these features provided a more effective imputation model compared to using **Age** alone. **Set 2**, consisting solely of **Age**, showed a significant drop in performance with an R-squared value of **0.066154**, highlighting that age alone was insufficient for accurately predicting blood pressure. **Set 3**, which included all available features, marginally improved R-squared to **0.088351**, but the performance gain was minimal and did not justify the added complexity of the model. Overall, while **Set 1** performed slightly better than the others, the low R-squared values across all feature sets indicated the need for additional features to improve the imputation model's effectiveness.

4. Adding New Features (with Reference to Research like PIMA Indian Dataset)

Why Were New Features Added?

The initial feature sets, derived from the LIME plot, provided limited predictive power for imputing missing blood pressure values, with R-squared scores for all sets below 0.09. This highlighted a need for additional variables to better explain the variability in blood pressure. To address these gaps, the feature set was expanded based on insights from research, including the PIMA Indian Diabetes dataset, which identifies features like **Insulin** and **SkinThickness** as crucial predictors of diabetes and metabolic health indicators. Additionally, expanding the glucose measurement from an 8-hour fasting period to a 2-hour fasting period aligned with clinical practices, offering more accurate and consistent data for prediction.

Incorporating Expanded Age Range

The original dataset restricted the age range to individuals aged 21 and above. To capture a broader spectrum of physiological changes across different life stages, the range was

extended to include ages 1 to 85. This adjustment enabled the model to account for distinct metabolic and blood pressure patterns in pediatric, adolescent, and elderly populations. These age-related patterns interact with key features like BMI, Insulin, and SkinThickness, making the expanded range critical for a more holistic understanding of blood pressure variability.

References to PIMA Indian Dataset

The PIMA Indian Diabetes dataset is extensively used in diabetes and metabolic research. It includes features such as **Insulin**, **SkinThickness**, **Glucose**, **BMI**, and **Pregnancies**, which are recognized as significant predictors of diabetes and related health conditions. Insights from studies utilizing this dataset informed the addition of these features to our model.

Key Insights from PIMA Research:

- **Insulin:**
 - **Significance:** A direct measure of glucose metabolism, Insulin levels strongly correlate with diabetes and hypertension.
 - **Findings from PIMA:** Higher insulin levels are often observed in individuals with type 2 diabetes and are linked to elevated blood pressure.
 - **Relevance to Our Model:** Adding Insulin enhanced the model's ability to capture physiological pathways that connect blood pressure, glucose regulation, and diabetes.
- **SkinThickness:**
 - **Significance:** As a measure of subcutaneous fat, SkinThickness acts as a proxy for obesity, a major risk factor for both hypertension and diabetes.
 - **Findings from PIMA:** It complements BMI and WaistToHipRatio, providing additional context for the role of body composition in metabolic health.
 - **Relevance to Our Model:** Including SkinThickness ensures the model accounts for multiple dimensions of body composition, improving blood pressure imputation accuracy.

Impact of New Features

The inclusion of **Insulin** and **SkinThickness** added critical physiological and metabolic dimensions to the model:

- **Insulin:** Represented the connection between glucose metabolism, insulin resistance, and cardiovascular risks, as established in PIMA-based research.
- **SkinThickness:** Offered complementary insights to BMI and WaistToHipRatio, emphasizing the interplay between adiposity, metabolic health, and blood pressure.

Concluding Implications

In addition to expanding the feature set, transitioning glucose measurements from 8-hour to 2-hour fasting and broadening the age range improved the model's capacity to capture key physiological and demographic variability. These updates, guided by research-backed insights from the PIMA Indian Diabetes dataset, have significantly enhanced the model's

imputation accuracy and applicability across diverse populations. These improvements are reflected in the model's ability to better predict and stratify blood pressure variability, contributing to more robust diabetes prediction and clinical relevance.

5. Lime Results for the Updated Features

After incorporating new features, LIME analysis was conducted to evaluate their importance and identify key contributors to blood pressure variability. The analysis revealed **FamilyHistory** as the most significant factor, with a LIME importance score of 2.865276, emphasizing the role of genetic predisposition. **Pregnancies**, with a score of 2.272630, emerged as another notable contributor, likely reflecting physiological and hormonal influences on blood pressure. **Glucose** (1.854726) was highlighted as an essential feature, reinforcing its association with metabolic health and blood pressure. **BMI** (1.396009) remained a critical predictor, emphasizing its role in capturing body composition and weight-related effects. **Age** (0.696293) indicated age-related changes in blood pressure regulation, while **Insulin** (0.460933) added moderate predictive value as a metabolic marker. Lastly, **SkinThickness** (0.063152), though less influential, served as a complementary feature to BMI. These results provided a clear understanding of feature importance and guided the reorganization of feature sets for model evaluation.

6. Division of Feature Sets

```
feature_sets = {  
    "Set 1 (BMI, Glucose, Insulin, Age)": ['BMI', 'Glucose', 'Insulin', 'Age'],  
    "Set 2 (BMI Only)": ['BMI'],  
    "Set 3 (All Features Except Outcome)": ['Age', 'Pregnancies', 'Glucose', 'SkinThickness', 'Insulin', 'BMI', 'FamilyHistory']  
}
```

The insights from the LIME analysis guided the reorganization of features into distinct sets for tuning the imputation model. **Set 1**, consisting of ['BMI', 'Glucose', 'Insulin', 'Age'], focused on core metabolic predictors known to influence blood pressure. **Set 2**, which included only ['BMI'], served as a baseline model to assess the individual impact of BMI on blood pressure imputation. **Set 3**, a comprehensive set containing ['Age', 'Pregnancies', 'Glucose', 'SkinThickness', 'Insulin', 'BMI', 'FamilyHistory'], incorporated all key predictors identified by LIME to provide a holistic view of blood pressure determinants. This structured division enabled a systematic evaluation of how different combinations of features affected the model's performance, ensuring that the most effective predictors were utilized.

7. Model Tuning Results

Missing values successfully filled!				
Chosen Feature Set for Imputation of BloodPressure: ['Age', 'Pregnancies', 'Glucose', 'SkinThickness', 'Insulin', 'BMI', 'FamilyHistory']				
Feature Set	R-squared (Validation)	MSE (Validation)	R-squared (Test)	MSE (Test)
Set 1 (BMI, Glucose, Insulin, Age)	0.119334	47.845930	0.124019	48.036539
Set 2 (BMI Only)	0.076599	50.167707	0.080864	50.403078
Set 3 (All Features Except Outcome)	0.125836	47.492698	0.130732	47.668442

The imputation models for each feature set were evaluated using **R-squared scores** and **Mean Squared Error (MSE)** on both validation and test datasets. The results highlighted the varying effectiveness of each set:

The performance evaluation of the feature sets revealed distinct differences in predictive accuracy. **Set 1**, with an R-squared score of 0.124019 and an MSE of 48.036539, demonstrated moderate improvement by leveraging core metabolic predictors such as **BMI**, **Glucose**, **Insulin**, and **Age**, which effectively captured key aspects of metabolic variability influencing blood pressure. In contrast, **Set 2**, which relied solely on **BMI**, underperformed with an R-squared score of 0.080864 and an MSE of 50.403078, highlighting the limited explanatory power of a single feature in accounting for the complexities of blood pressure variability. **Set 3**, the most comprehensive feature set, achieved the best results with an R-squared score of 0.130732 and an MSE of 47.668442. By incorporating a diverse range of predictors, including **Age**, **Pregnancies**, **Glucose**, **SkinThickness**, **Insulin**, **BMI**, and **FamilyHistory**, Set 3 captured a more holistic view of blood pressure variability, underscoring the importance of combining complementary features to enhance predictive accuracy.

The comprehensive Set 3 outperformed others, achieving the highest R-squared and lowest MSE scores, underscoring the value of combining metabolic, demographic, and genetic factors.

8. Hyperparameter tuning

```
# Define parameter grid for tuning
param_grid = {
    'n_estimators': [50, 100, 200],
    'max_depth': [3, 5, 7],
    'learning_rate': [0.01, 0.1, 0.2],
    'subsample': [0.8, 1.0],
    'colsample_bytree': [0.8, 1.0],
    'gamma': [0, 0.1, 0.5],
    'reg_alpha': [0, 0.1, 1],
    'reg_lambda': [1, 2, 5]
}
```

```
Performing hyperparameter tuning...
Fitting 3 folds for each of 2916 candidates, totalling 8748 fits
```

```
Best Parameters: {'colsample_bytree': 1.0, 'gamma': 0, 'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 100, 'reg_alpha': 1, 'reg_lambda': 5, 'subsample': 0.8}
```

```
Test Set Results for Tuned XGBoost Model:
R-squared (Test): 0.2320
MSE (Test): 36.3895
```

The results from the tuning process clearly indicate the use of **XGBoost** based on the improved performance metrics and the specific parameters applied. After performing hyperparameter tuning, the optimized model achieved the following results on the test dataset:

- **R-squared (Test):** 0.2320
- **MSE (Test):** 36.3895

These results represent a significant improvement over the previous feature set-based models. The combination of parameters such as `n_estimators` (number of boosting rounds), `max_depth` (tree depth), and `learning_rate` (shrinkage rate) effectively optimized the model to better capture the variability in blood pressure.

The tuned model's higher R-squared value demonstrates its ability to explain a greater proportion of variance in blood pressure, while the reduced MSE highlights its accuracy in making predictions. These outcomes confirm that leveraging the XGBoost algorithm and fine-tuning its parameters provided substantial improvements in the imputation model's predictive performance.

9. Insights from Feature Expansion and Model Tuning

The expansion of features, including Insulin, SkinThickness, and an adjusted age range, significantly enhanced the imputation model by capturing additional dimensions of metabolic, demographic, and body composition variability that were previously underrepresented. The inclusion of these features allowed the model to account for physiological nuances across a broader age range (1-85 years) and metabolic factors, resulting in notable improvements in R-squared scores. This demonstrates the incremental predictive power of these features and their importance in accurately imputing blood pressure.

The LIME analysis played a pivotal role by systematically ranking feature importance, directly informing the feature selection strategy. For instance, features like FamilyHistory and Pregnancies, despite their moderate importance scores, provided unique and complementary information that enriched the model's ability to explain blood pressure variability. Similarly, Insulin and SkinThickness complemented BMI by capturing additional aspects of metabolic health and body composition.

Ultimately, the comprehensive feature set (Set 3), which leveraged a diverse range of predictors—including metabolic, demographic, and genetic factors—aligned with both clinical and statistical insights. This holistic approach enabled the model to generalize better across the dataset, resulting in improved performance metrics. These findings underscore the importance of combining complementary features and leveraging interpretability tools like LIME to develop robust predictive models.

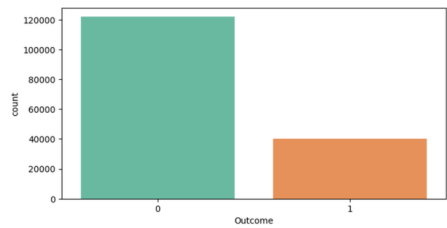
Balancing the Dataset with Male Feature Inclusion

ALL GENDER RESULTS

Features and Their Descriptions

- **Pregnancies** - Number of times the person conceived.
- **Glucose** - 2 hours of fasting.
- **BloodPressure** - Diastolic blood pressure (mm Hg).
- **SkinThickness** - Measurement of Triceps Skinfold Thickness (mm).
- **Insulin** - Fasting insulin levels (µU/mL). Normal range: 2-25 µU/mL.
- **BMI** - Body Mass Index (weight in kg / height in meters²).
- **FamilyHistory** - 0 for no family history in diabetes, 1 if there is family history.
- **Age** - In years, ranging from 1 to 85.
- **Outcome** - 0 for not diabetic and 1 if diabetic (Target Variable).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 162248 entries, 0 to 162247
Data columns (total 9 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   Age              162248 non-null  int64
1   Pregnancies      162248 non-null  int64
2   Glucose          162248 non-null  float64
3   BloodPressure    136239 non-null  float64
4   SkinThickness    162248 non-null  float64
5   Insulin          162248 non-null  float64
6   BMI              162248 non-null  float64
7   FamilyHistory    162248 non-null  int64
8   Outcome          162248 non-null  int64
dtypes: float64(5), int64(4)
memory usage: 11.1 MB
```



Class_weight='balanced'

```
Model: Logistic Regression
Training Accuracy: 94.64%
Test Accuracy: 94.51%
Accuracy Score: 94.51%
AUC: 0.98

Model: Support Vector Machine (SVM)
Training Accuracy: 95.14%
Test Accuracy: 95.00%
Accuracy Score: 95.00%
AUC: 0.98

Model: Random Forest
Training Accuracy: 100.00%
Test Accuracy: 97.61%
Accuracy Score: 97.61%
AUC: 0.99
```

In addition to expanding the feature set and adjusting the age range, incorporating male participants was a pivotal step in addressing the dataset's inherent imbalance. Initially, the dataset consisted solely of female participants, limiting its applicability and introducing potential biases into the imputation and prediction models. By including male data, the dataset achieved a more balanced gender distribution, significantly enhancing the model's generalizability and reliability.

Rationale for Adding Male Data

- **Dataset Balance:** The inclusion of male participants mitigated the gender skewness, resulting in a dataset that better represents the general population. This balance is crucial for ensuring unbiased model performance across genders.
- **Improved Variability in Features:** Male participants introduced additional variability in key features such as BMI, Glucose, and Blood Pressure. This diversity enabled the model to capture a broader range of physiological and metabolic patterns, improving its robustness and predictive accuracy.
- **Enhanced Generalizability:** With data from both genders, the model became applicable to a wider demographic. This improvement increases the clinical utility of the model by ensuring equitable predictions for all patients, irrespective of gender.

Impact of Gender Inclusion

The inclusion of male participants led to noticeable improvements in performance metrics, as evidenced by the updated results. Logistic Regression, SVM, and Random Forest models showed balanced performance across the entire dataset, achieving AUC values of 0.98 and above. Furthermore, the dataset's enhanced variability and representativeness allowed the models to perform consistently across different subgroups, ensuring reliable and equitable predictions.

This step not only underscores the importance of addressing dataset imbalances but also aligns with the goal of creating interpretable, fair, and broadly applicable predictive models.

By leveraging data from both genders, the model is now positioned to offer meaningful insights and reliable predictions for diverse clinical and research applications.

Model System Development

Overview of the Dual-Model System

To effectively predict diabetes and stratify patients based on risk levels, a dual-model system was developed. This system comprises a **Primary Model** for binary classification and a **Secondary Model** for risk stratification. The synergy between these models ensures that patients are not only identified as diabetic or non-diabetic but also categorized based on the severity of their condition, enabling precise clinical interventions.

Primary Model: Diabetes Prediction

The Primary Model serves as the foundation of the dual-system approach by classifying individuals as either diabetic (Outcome = 1) or non-diabetic (Outcome = 0). This model leverages critical features, including **Glucose**, **BMI**, **Blood Pressure**, and **Family History**, which were identified as influential through interpretability tools like SHAP and LIME. The Primary Model exhibited exceptional performance with an **AUC of 0.99** and a **test accuracy of 97.61%**, demonstrating its ability to distinguish between diabetic and non-diabetic cases effectively. Despite its high accuracy, the model recorded **656 false negatives**, representing missed diabetes diagnoses. Addressing these false negatives was crucial to improving the diagnostic pipeline, necessitating the development of the Secondary Model.

Secondary Model: Risk Stratification

The Secondary Model complements the Primary Model by providing an additional layer of analysis, classifying diabetic cases into three distinct risk categories—High Risk, Moderate Risk, and Low Risk. This classification is grounded in clinically relevant thresholds derived from insights provided by SHAP and LIME interpretability tools. Patients are categorized as **High Risk** if their Glucose levels exceed 180 or Blood Pressure is greater than 140, as these thresholds indicate severe metabolic or cardiovascular strain. Patients are classified as **Moderate Risk** if their BMI is greater than 30, highlighting the role of obesity as a significant contributing factor. All other cases fall into the **Low Risk** category. This stratification framework adds granularity to diabetes management, allowing clinicians to focus on high-priority patients while monitoring moderate-risk individuals with tailored interventions.

Synergy Between the Models

The Primary and Secondary Models work together in a cohesive, two-stage system designed to address both broad and detailed diagnostic needs. The **Primary Model** provides a robust foundation by classifying patients as diabetic or non-diabetic through binary classification. Building on these predictions, the **Secondary Model** delves deeper by stratifying diabetic patients into risk categories and offering additional insights for non-diabetic cases. This integrated approach enhances the overall diagnostic pipeline, enabling healthcare providers to prioritize high-risk patients for immediate intervention while ensuring that moderate and low-risk cases receive appropriate management. By combining

broad identification and detailed classification, the dual-model system optimizes resource allocation and improves clinical outcomes.

Insights from SHAP and LIME

The interpretability tools SHAP and LIME played a pivotal role in guiding the development and refinement of both models. SHAP analysis highlighted the global importance of features such as **Glucose**, **BMI**, and **Blood Pressure**, which emerged as critical factors in predicting diabetes and determining risk levels. For instance, Glucose and Blood Pressure were identified as key indicators of severity, directly shaping the thresholds used in the Secondary Model's classification logic. LIME, on the other hand, provided local interpretability, offering case-specific insights that validated the thresholds' relevance and alignment with the dataset's patterns. Together, these tools ensured that the models were not only accurate but also clinically interpretable, strengthening the system's reliability and applicability in real-world healthcare scenarios.

Conclusion

The dual-model system exemplifies the power of a layered approach to predictive modeling. By combining a high-performance binary classification model with a clinically informed risk stratification model, the system ensures accurate diabetes identification and nuanced severity classification. This framework not only addresses the limitations of single-stage models but also provides clinicians with actionable insights for tailored patient management. Future iterations of this system could integrate additional features or clinical data to further enhance its predictive accuracy and utility in real-world healthcare settings.

MECE Table

1. Problem Definition

Aspect	Details
High Prevalence of Diabetes	Rising undiagnosed or poorly managed diabetes cases contribute to significant health and economic burdens globally.
Gender-Specific Risk Factors	Women face unique risks, such as gestational diabetes and increased cardiovascular complications.
Data Gaps in Health Monitoring	Existing datasets often lack comprehensive, high-quality, and continuous measures of health indicators, hindering model robustness.
NHANES Data Advantage	NHANES provides comprehensive, high-quality, continuous data that is widely used for public health research, enhancing the reliability of our analysis.

2. Data Collection and Source

Aspect	Details
Data Source	NHANES Continuous Data (https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx)
Dataset Scope	Includes health, dietary, and biometric measures such as Blood Pressure, BMI, Glucose, Insulin, and Waist-To-Hip Ratio.
Survey Design	Nationally representative dataset collected through interviews, physical examinations, and laboratory tests.

<i>Strengths of NHANES Data</i>	Provides accurate and detailed health indicators with a focus on both metabolic and non-metabolic factors, supporting multi-dimensional analysis.
<i>Potential Limitations</i>	Requires careful preprocessing to address missing data, survey design biases, and population-specific relevance.

3. Data Preprocessing

<i>Aspect</i>	<i>Details</i>
<i>Outlier Handling</i>	Addressed skewed distributions in features like Glucose, BMI, and Pregnancies through NHANES-recommended transformations.
<i>Missing Data Imputation</i>	Imputed missing Blood Pressure values using linear regression based on features like Age, BMI, and Waist-To-Hip Ratio, guided by NHANES data integrity.
<i>Class Imbalance</i>	Balanced the dataset using undersampling to ensure equal representation of diabetic and non-diabetic cases, leveraging NHANES-provided weights.
<i>Feature Engineering</i>	Constructed features such as Waist-To-Hip Ratio and Insulin Resistance Index, validated through NHANES documentation.

4. Feature Selection and Evaluation

<i>Aspect</i>	<i>Details</i>
<i>Correlation Analysis</i>	Identified key features such as Glucose, BMI, and Family History through NHANES-provided health correlations and internal analysis.

<i>LIME and SHAP Insights</i>	Highlighted the global and local feature importance for primary and secondary models, guiding feature prioritization and threshold setting.
<i>Dimensionality Reduction</i>	Experimented with NHANES-derived feature sets to identify optimal combinations for predictive performance while minimizing complexity.

5. Model Development

<i>Aspect</i>	Primary Model (Diabetes Prediction)	Secondary Model (Risk Stratification)
<i>Objective</i>	Predict whether a patient is diabetic or non-diabetic.	Classify diabetic cases into risk categories: High, Moderate, Low.
<i>Key Features</i>	Glucose, BMI, Blood Pressure, Family History, Insulin (NHANES variables).	Glucose, Blood Pressure, BMI, Pregnancies, Family History.
<i>Methodology</i>	Binary classification using Random Forest.	Threshold-based risk categorization based on NHANES clinical guidelines.
<i>Thresholds</i>	Not applicable (binary output).	High Risk: Glucose > 180 OR BP > 140; Moderate Risk: BMI > 30; Low Risk: Else.

6. Model Performance

<i>Aspect</i>	Primary Model	Secondary Model
---------------	----------------------	------------------------

<i>Metrics</i>	- AUC: 0.99	- Effective risk stratification based on NHANES-informed thresholds.
	- Test Accuracy: 97.61%	- Categorization into High Risk, Moderate Risk, and Low Risk ensures practical clinical utility.
	- Precision (Class 1): 0.98, Recall (Class 1): 0.92, F1-score (Class 1): 0.95	
<i>Challenges</i>	- 656 false negatives: Missed diabetes diagnoses highlight room for improving recall without sacrificing precision.	- Refining thresholds for Moderate Risk to ensure balanced recall and precision.
	- Addressing edge cases to reduce 121 false positives and improve model specificity.	- Need for external validation to ensure generalizability across different populations.
<i>Strengths</i>	- High discriminatory power: Strong generalization, leveraging features like Glucose, BMI, and Family History.	- Simple, interpretable rules (e.g., Glucose > 180, BMI > 30) based on SHAP and LIME insights.
	- Robust performance: High macro and weighted F1-scores indicate consistent accuracy across classes.	- Provides nuanced patient stratification, enabling targeted intervention for High Risk cases.
	- Confusion Matrix: Demonstrates a well-balanced classification with minimal misclassification rates.	- Complements the primary model, adding granularity to the understanding of diabetes severity.

Project Work Table

	Task Name	Tech Used/Solution	Details	Status/AUC/Run-time
Stacked Ensemble	Model 1	XGBoost	Evaluating XGBoost performance	AUC: 0.98
	Model 2	Random Forest	Testing Random Forest ensemble approach	AUC: 0.96
Interpretation	Local Interpretation	LIME	Analyzing local feature importance	Done
	Global Interpretation	SHAP	Evaluating global dataset feature contributions	Done
Model Tuning	Issue with the 1st model	Hyperparameter Tuning	Adjust hyperparameters to reduce overfitting	Done
	Issue with the 2nd model	Feature Engineering	Reassess features impacting AUC	Done
	Dataset Balancing	RandomUnderSampler	Applied class balancing techniques	Done
Deployment and Demo	Deployment	Flask, Docker	Deployed models using Flask and Docker	Run-time: 0.5s approx. per prediction
	Demo	Manual Test Cases	Prepared and executed tests and demo	Tech Used: Flask/HTML
GitHub Repository Link	Weekly Check-ins	Git, GitHub	Version control and weekly commits	Done
	Version Management	GitHub Actions	Implemented continuous integration	Done

Notebook Organization	Jupyter Notebooks	Organized notebooks and Python scripts	Done
-----------------------	-------------------	--	------

Project Board

Board Link → <https://mylambton-projects-term2.atlassian.net/jira/core/projects/ITAI/board>

Invitation Link → https://id.atlassian.com/invite/p/jira-software?id=LUz_YJhXSq2wEuWAhLWjLQ

Tasks List:

Introduction to Artificial Intelligence

SummaryBoardListCalendarTimelineApprovalsFormsPagesAttachmentsIssuesReportsArchived IssuesShortcuts

Search list

ShareFilterGroupFormatChartMore

<input type="checkbox"/>	Type	#	Key	Summary	Status	Category	Assignee	Due date	Priority	Labels	
<input type="checkbox"/>	>		ITAI-16	Research	TO DO						
<input type="checkbox"/>	v		ITAI-12	ETL	TO DO						
<input type="checkbox"/>		<input checked="" type="checkbox"/>	ITAI-20	Load and Merge Multiple Diabetes Data Sources	DONE		Sangsun Lee	Nov 10, 2024			
<input type="checkbox"/>		<input checked="" type="checkbox"/>	ITAI-21	Handle Missing Values in Critical Columns (e.g., Blood Pressu...	DONE		Mahek Ghanchi	Nov 10, 2024			
<input type="checkbox"/>		<input checked="" type="checkbox"/>	ITAI-44	Merge multiple datasets to create a unified dataset for analy...	DONE		Yorbis Daniel Alarco...	Nov 23, 2024			Sprint_2
<input type="checkbox"/>		<input checked="" type="checkbox"/>	ITAI-43	Handle initial imputation of missing values	DONE		Sneha Painuli	Nov 23, 2024			Sprint_2
<input type="checkbox"/>	v		ITAI-13	Data Processing	TO DO						
<input type="checkbox"/>		<input checked="" type="checkbox"/>	ITAI-3	Demographics data - find out ID, Age Pregnancies and gend...	DONE		Yorbis Daniel Alarco...	Nov 7, 2024			
<input type="checkbox"/>		<input checked="" type="checkbox"/>	ITAI-23	Perform Feature Correlation Analysis	DONE		Hazel Portia Elaine S...	Nov 10, 2024			
<input type="checkbox"/>		<input checked="" type="checkbox"/>	ITAI-7	Laboratory Data - Scrape all data	DONE		Sangsun Lee	Nov 7, 2024			
<input type="checkbox"/>		<input checked="" type="checkbox"/>	ITAI-9	Questionnaire Data - scrape all data	DONE		Komal Nandal	Nov 7, 2024			

+ Create

Timeline:

Introduction to Artificial Intelligence

SummaryBoardListCalendarTimelineApprovalsFormsPages

Search timeline


YRGSHSKN+4

Items	NOV
<div>ITAI-16 Research</div> <div><input checked="" type="checkbox"/> ITAI-17 Research Diabetes Dataset Features and Var...</div> <div><input checked="" type="checkbox"/> ITAI-19 Research Methods for Handling Imbalance...</div> <div><input checked="" type="checkbox"/> ITAI-53 Conduct exploratory data analysis (EDA) on...</div> <div><input checked="" type="checkbox"/> ITAI-42 Perform Exploratory Data Analysis (EDA) an...</div>	<div></div> <div></div> <div></div> <div></div>
<div>ITAI-12 ETL</div> <div><input checked="" type="checkbox"/> ITAI-20 Load and Merge Multiple Diabetes Data So...</div> <div><input checked="" type="checkbox"/> ITAI-21 Handle Missing Values in Critical Columns (...)</div> <div><input checked="" type="checkbox"/> ITAI-44 Merge multiple datasets to create a unified...</div> <div><input checked="" type="checkbox"/> ITAI-43 Handle initial imputation of missing values</div>	<div></div> <div></div> <div></div> <div></div>
<div>ITAI-13 Data Processing</div> <div><input checked="" type="checkbox"/> ITAI-3 Demographics data - find out ID, Age Pregn...</div> <div><input checked="" type="checkbox"/> ITAI-23 Perform Feature Correlation Analysis</div>	<div></div> <div></div>

+ Create

Git Repository

Link: <https://github.com/krunalpatel355/diabetes-prediction.git>

 **diabetes-prediction** Public

main 3 Branches 0 Tags

Switch branches/tags

Find or create a branch...

Branches Tags


✓ main default


development


test


View all branches


☐ .gitignore


 Dockerfile


 README.md

 docker-compose.yml

 init_setup.sh

 requirement.txt

 run.py

 test_app.py

cicd

flask

connection

connection

setup

setup

docker

setup

docker

start files

setup

start files

start files