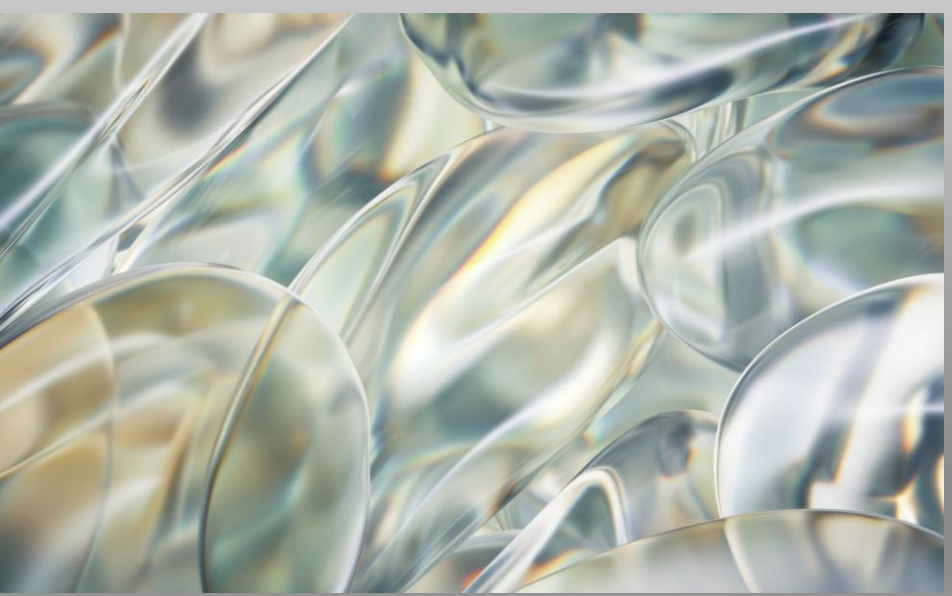
Title:

From Symptoms to Solutions: Bayesian Logistic and Linear Regression Models for Diabetes Prediction



Group Members:
Sheril Sarwar 238063
Sourav Sarker 237854
Technical University of Dortmund
30th January, 2025.

Introduction and Dataset Overview

1. This project aims to predict diabetes risk using **Bayesian Logistic Regression** for classification and **Bayesian Linear Regression** for continuous probability estimation, leveraging statistical insights for clinical decision-making. It focuses on analyzing key symptoms such as **Polyuria**, **Polydipsia**, and **Age** to identify significant predictors of diabetes.

2. The dataset consists of 520 observations and 17 features, including clinical symptoms (e.g., **Polyuria**, **Polydipsia**) and the diabetes classification (`Positive` or `Negative`).

Dataset Souce: https://www.kaggle.com/datasets/yasserhessein/early-stage-diabetes-risk-prediction-dataset/data

Research Goals and Analytical Approach

Our primary inquiry in this study focuses on understanding how clinical features influence diabetes risk. Specifically, we ask:

How do predictors such as Age, Polyuria, and Polydipsia affect diabetes diagnosis, and how can Bayesian Logistic and Linear Regression improve predictive performance and interpretability?

Objectives

Objective 1

To quantitatively assess the impact of individual predictors such as **Age**, **Polyuria**, and **Polydipsia** on diabetes diagnosis using **Bayesian Logistic Regression**.

Objective 2

To explore relationships between predictors and the continuous probabilities of diabetes diagnosis using **Bayesian Linear Regression**.

Objective 3

To enhance the predictive accuracy and reliability of the model by incorporating **Bayesian Inference** for uncertainty quantification and prior knowledge integration.

Objective 4

To compare and validate model performance using **cross-validation techniques (e.g., K-Fold)** and statistical criteria like **AIC** and **BIC** for robust evaluation.

Key Variables

Clinical Features:

- **Age**: Significant predictor; risk increases with age.
- **Polyuria**: Binary variable indicating excessive urination.
- **Polydipsia**: Binary variable indicating excessive thirst.

Statistical Outputs:

- **Predicted Probability**: Likelihood of diabetes from Bayesian Logistic Regression.
- **Class**: Target variable (`Positive`/`Negative`) for diabetes diagnosis.

Evaluation Metrics:

- **Accuracy**: Overall classification performance.
- **F1-Score**: Balance between precision and recall.
- **Precision/Recall**: Measures of true positives and sensitivity.

The key insights of the dataset

Age

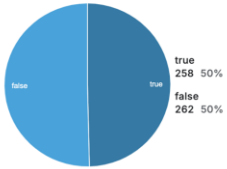
Valid	520	100%
Mismatched	0	0%
Missing	0	0%
Mean	48	
Std. Deviation	12.1	
Quantiles	16	Min
	39	25%
	48	50%
	57	75%
	90	Max

Gender

Male	63%
Female	37%

Valid	520	100%
Mismatched	0	0%
Missing	0	0%
Unique	2	
Most Common	Male	63%

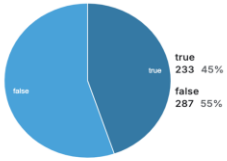
Polyuria



true 258 50%
false 262 50%

Valid	520	100%
Mismatched	0	0%
Missing	0	0%
True	258	50%
False	262	50%

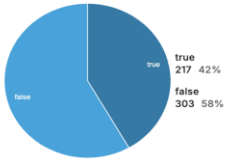
Polydipsia



true 233 45%
false 287 55%

Valid	520	100%
Mismatched	0	0%
Missing	0	0%
True	233	45%
False	287	55%

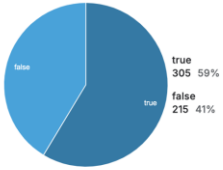
sudden weight loss



true 217 42%
false 303 58%

Valid	520	100%
Mismatched	0	0%
Missing	0	0%
True	217	42%
False	303	58%

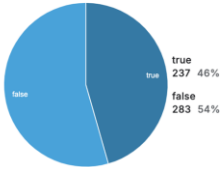
weakness



true 305 59%
false 215 41%

Valid	520	100%
Mismatched	0	0%
Missing	0	0%
True	305	59%
False	215	41%

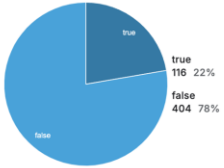
Polyphagia



true 237 46%
false 283 54%

Valid	520	100%
Mismatched	0	0%
Missing	0	0%
True	237	46%
False	283	54%

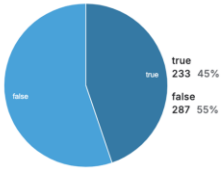
Genital thrush



true 116 22%
false 404 78%

Valid	520	100%
Mismatched	0	0%
Missing	0	0%
True	116	22%
False	404	78%

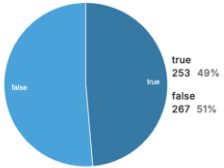
visual blurring



true 233 45%
false 287 55%

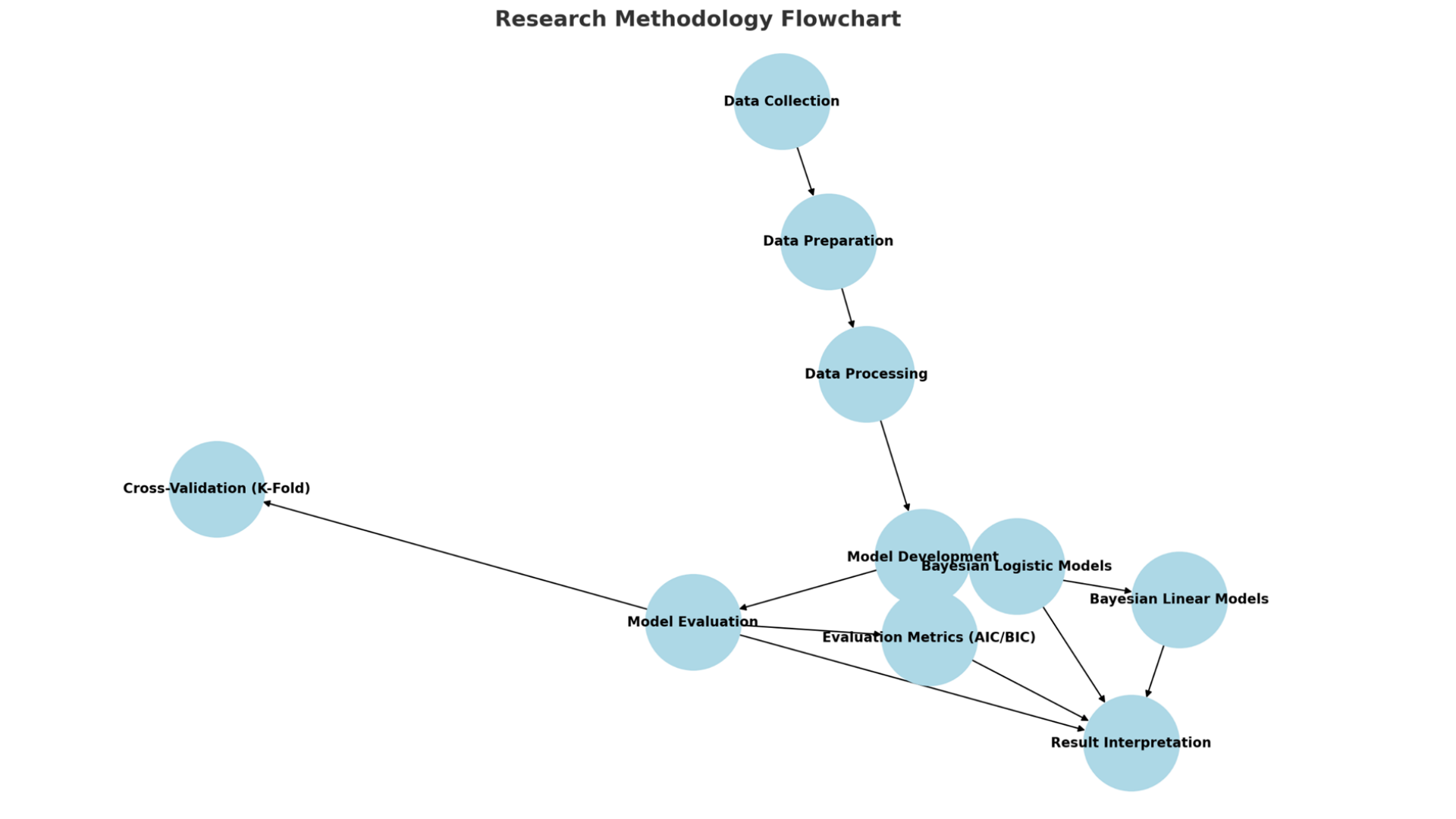
Valid	520	100%
Mismatched	0	0%
Missing	0	0%
True	233	45%
False	287	55%

Itching



true 253 49%
false 267 51%

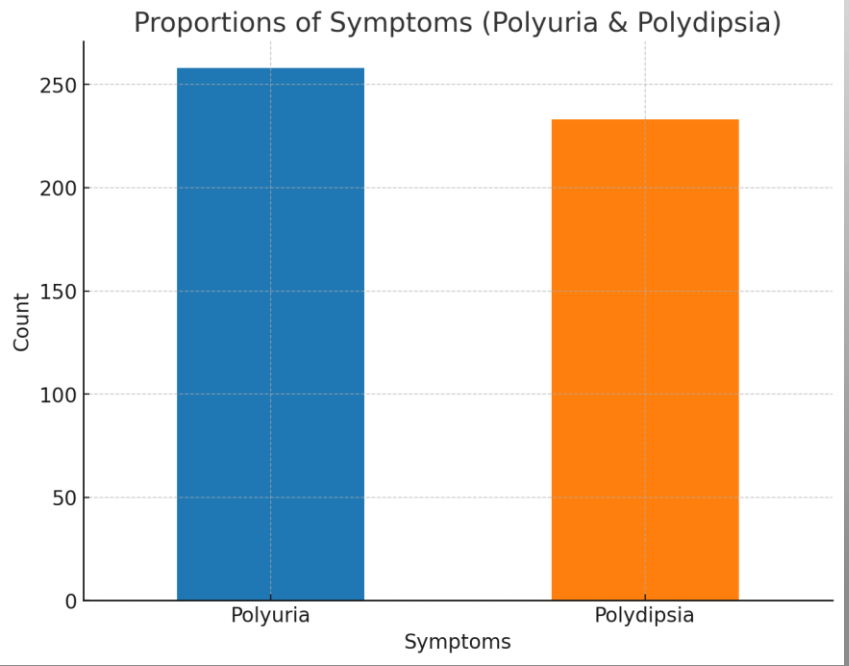
Valid	520	100%
Mismatched	0	0%
Missing	0	0%
True	253	49%
False	267	51%

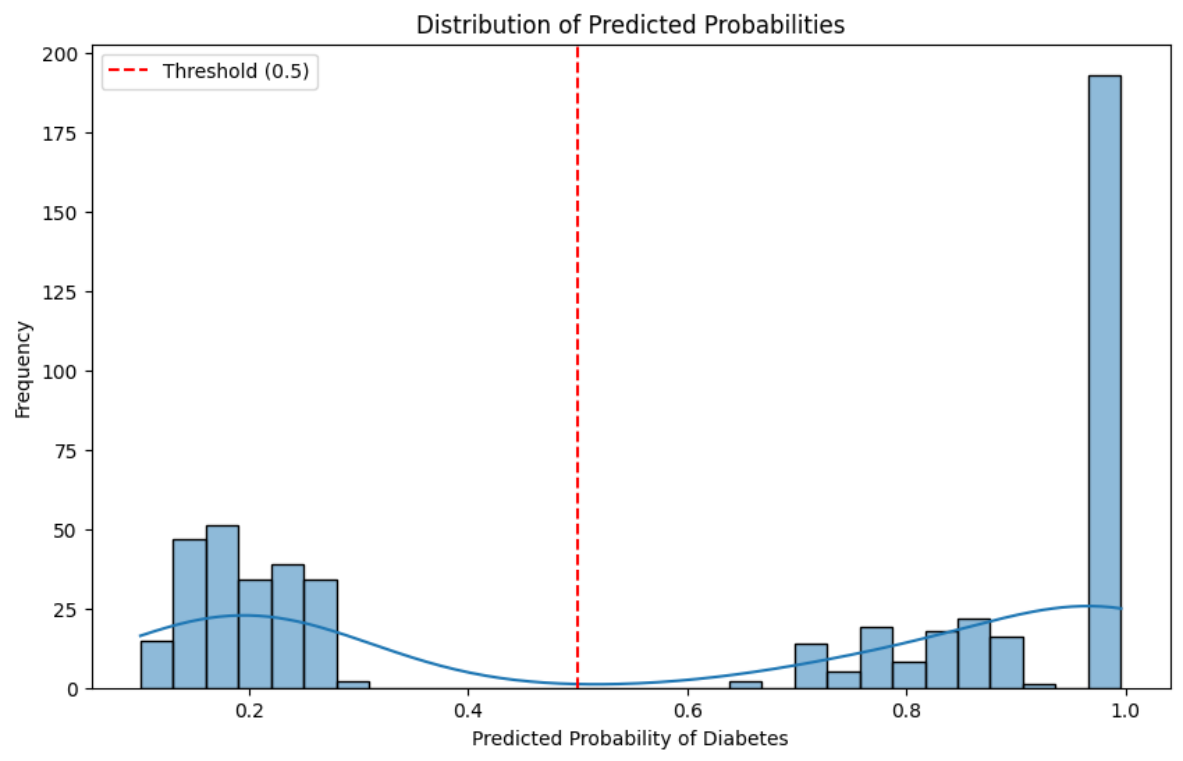

Research Methodology Flowchart
Data Collection
Data Preparation
Data Processing
Model Development
Bayesian Logistic Models
Bayesian Linear Models
Model Evaluation
Evaluation Metrics (AIC/BIC)
Cross-Validation (K-Fold)
Result Interpretation

Data Preprocessing

- Converted categorical variables (e.g., Polyuria, Polydipsia) to binary numeric values.
- Removed samples with missing values.
- Ensured balanced train-test split for reliable evaluation.

Visualization




Distribution of Predicted Probabilities
Threshold (0.5)
Frequency
Predicted Probability of Diabetes
0
25
50
75
100
125
150
175
200
0.2
0.4
0.6
0.8
1.0

Model Descriptions

Bayesian Logistic Regression

Bayesian Logistic Regression is used to classify individuals as diabetes-positive or negative based on predictors such as **Age**, **Polyuria**, and **Polydipsia**. This model estimates the probability of diabetes while incorporating prior knowledge to enhance prediction accuracy and handle uncertainty. It provides interpretable coefficients to quantify the impact of each predictor on the likelihood of diabetes. Evaluation metrics such as **Accuracy**, **Precision**, **Recall**, and the **F1-Score** ensure the model's effectiveness.

Bayesian Linear Regression

Bayesian Linear Regression predicts continuous probabilities of diabetes using outputs from the logistic regression model. It models the linear relationship between predictors (e.g., **Age**, **Polyuria**) and diabetes probabilities while quantifying uncertainty in predictions. This makes it particularly useful for robust and reliable decision-making. The model’s performance is assessed using **Mean Squared Error (MSE)** and **R-Squared**, ensuring a strong fit and accurate predictions.

Optimization Output(Bayesian Logistic Regression)

- Optimization successfully terminated after 8 iterations.
- Final log-likelihood value: 0.319358.
- Model converged with statistically significant predictors.

Coefficients Interpretation

- Age: Negative coefficient, slight decrease in diabetes probability with age.
- Polyuria: Strong positive coefficient, increases diabetes odds by ~21.6x.
- Polydipsia: Strong positive coefficient, increases diabetes odds by ~28.5x.

Logistic Regression Accuracy

- Accuracy: 87%
- Bayesian Logistic Regression effectively classifies diabetes based on predictors.

Linear Regression Results

- Mean Squared Error (MSE): 0.01
- R-squared: 0.91
- Model predictions closely align with actual values.

Hyperparameter Tuning

- Grid Search used to optimize hyperparameters (C, solver).
- Best parameters: {'C': 0.1, 'solver': 'liblinear'}
- Regularization improved generalization.

Cross-Validation

- K-Fold Cross-Validation (5 splits) assessed model performance.
- Ensured robust evaluation across different data subsets.

Evaluation Metrics

- Accuracy: 87%
- Precision: 92%
- Recall: 86%
- F1 Score: 89%

Confusion Matrix

- True Positives: 57
- False Positives: 7
- True Negatives: 33
- False Negatives: 7

Model Comparison

Bayesian Logistic Regression:

- Accuracy: 87%
- Strengths: Predicts binary outcomes and provides uncertainty quantification.
- Key predictors: Age (-0.032), Polyuria (+3.19), Polydipsia (+3.44).

Bayesian Linear Regression:

- Mean Squared Error (MSE): 0.01
- R-squared: 0.91
- Strengths: Explains variance in probabilities and complements logistic regression.
- Predictors: Age, Polyuria, Polydipsia.

Bayesian Logistic Regression Results:

Logit Regression Results

Dep. Variable:	class	No. Observations:	416
Model:	Logit	Df Residuals:	412
Method:	MLE	Df Model:	3
Date:	Wed, 15 Jan 2025	Pseudo R-squ.:	0.5259
Time:	00:05:22	Log-Likelihood:	-132.85
converged:	True	LL-Null:	-280.21
Covariance Type:	nonrobust	LLR p-value:	1.380e-63

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0510	0.586	-0.087	0.931	-1.200	1.098
Age	-0.0322	0.013	-2.495	0.013	-0.058	-0.007
Polyuria	3.1936	0.409	7.806	0.000	2.392	3.996
Polydipsia	3.4378	0.498	6.902	0.000	2.462	4.414

Logistic Regression Accuracy: 0.87

Bayesian Linear Regression Results:

OLS Regression Results

Dep. Variable:	Predicted Probability	R-squared:	0.937
Model:	OLS	Adj. R-squared:	0.937
Method:	Least Squares	F-statistic:	2057.
Date:	Wed, 15 Jan 2025	Prob (F-statistic):	1.94e-247
Time:	00:05:22	Log-Likelihood:	393.14
No. Observations:	416	AIC:	-778.3
Df Residuals:	412	BIC:	-762.2
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]

Strengths

Accurate and interpretable predictions.

Robust evaluation with Bayesian inference.

Limitations

Sensitive to prior assumptions.

Computationally intensive for large datasets.

Practical Applications

1. **Early Detection of Diabetes**:
   The models can identify individuals at high risk of diabetes based on symptoms like Polyuria and Polydipsia, enabling early diagnosis and timely medical intervention. This helps reduce the progression of diabetes-related complications.

2. **Clinical Decision Support**:
   The probabilistic predictions from Bayesian models provide clinicians with interpretable insights, allowing them to make informed decisions about further diagnostic testing or preventive treatments.

3. **Personalized Risk Assessment**:
   By incorporating patient-specific data, the models can generate personalized diabetes risk profiles. This empowers healthcare providers to tailor treatment plans and prioritize high-risk individuals.

4. **Integration with Healthcare Systems**:
   These models can be integrated into electronic health record (EHR) systems to provide real-time risk predictions for diabetes during routine patient check-ups.

Take-Home Messages Project Link and QR code

- The analysis highlights key predictors of diabetes, with "Polyuria" and "Polydipsia" identified as the most significant symptoms, increasing the likelihood of diagnosis by over 20-fold. Logistic regression demonstrated strong performance, explaining 51.8% of the variance in diabetes risk (pseudo $R^2 = 0.5183$) and achieving an average cross-validated accuracy of 87%. Additionally, linear regression confirmed the reliability of the predicted probabilities, with an R^2 of 0.91. These findings emphasize the importance of symptom-based risk assessment for early detection and intervention.
- Project Link: https://www.kaggle.com/code/srvskr3245/from-symptoms-to-solutions-bayesian-logistic-and
- Email: sarwar.sheril@tu-dortmund.de
  sourav.sarker@tu-dortmund.de