



GlucoSense: AI-Powered Diabetes Detection for Early Intervention

Presented by: Charitha Duvvuru



Problem Statement:

Diabetes cases over the past fifteen years have bloomed all over the world. Lifestyle plays a very important role in it. In recent years, there has been an improvement in awareness regarding the health effects of diabetes. This has led to people getting themselves tested for diabetes than they would have earlier, as its risk can be reduced if it is predicted early.

The importance of early diabetes detection:



Early detection of diabetes is crucial to prevent severe complications such as cardiovascular diseases, kidney failure, and neuropathy.




Identifying the condition at an early stage enables timely interventions, including lifestyle changes and medical treatments, which can help manage blood sugar levels, reduce long-term health risks, and improve overall quality of life.



Early diagnosis also lowers healthcare costs by minimizing the need for extensive treatments and hospitalizations.

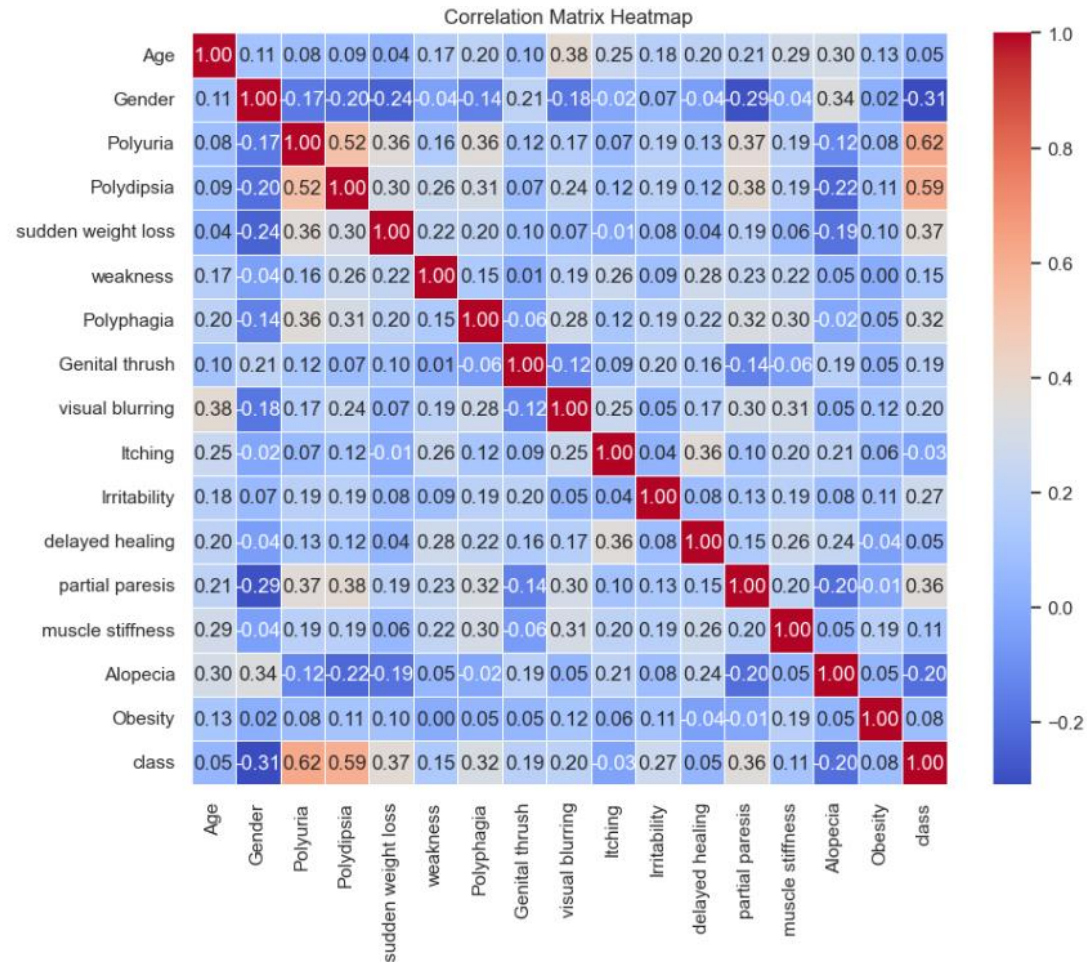
Data Collection:

- For the GlucoSense: AI-Powered Diabetes detection for early intervention project, the data used for analysis and model development was sourced from Kaggle, utilizing the "diabetes_risk_prediction_dataset.csv", which contains comprehensive healthcare and lifestyle statistics essential for predicting diabetes risk.
- Features in dataset are polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, class.



Data Exploration (EDA) and Data Preprocessing

- Checked dataset structure (520 rows, 17 columns) and ensured no missing values.
- Removed 269 duplicate rows for unbiased analysis.
- Identified key features like **Polyuria** and **Polydipsia** as strong predictors through statistical analysis and visualizations.
- Used a correlation matrix to explore relationships between variables.
- Cleaned data by removing duplicates and outliers (IQR method).

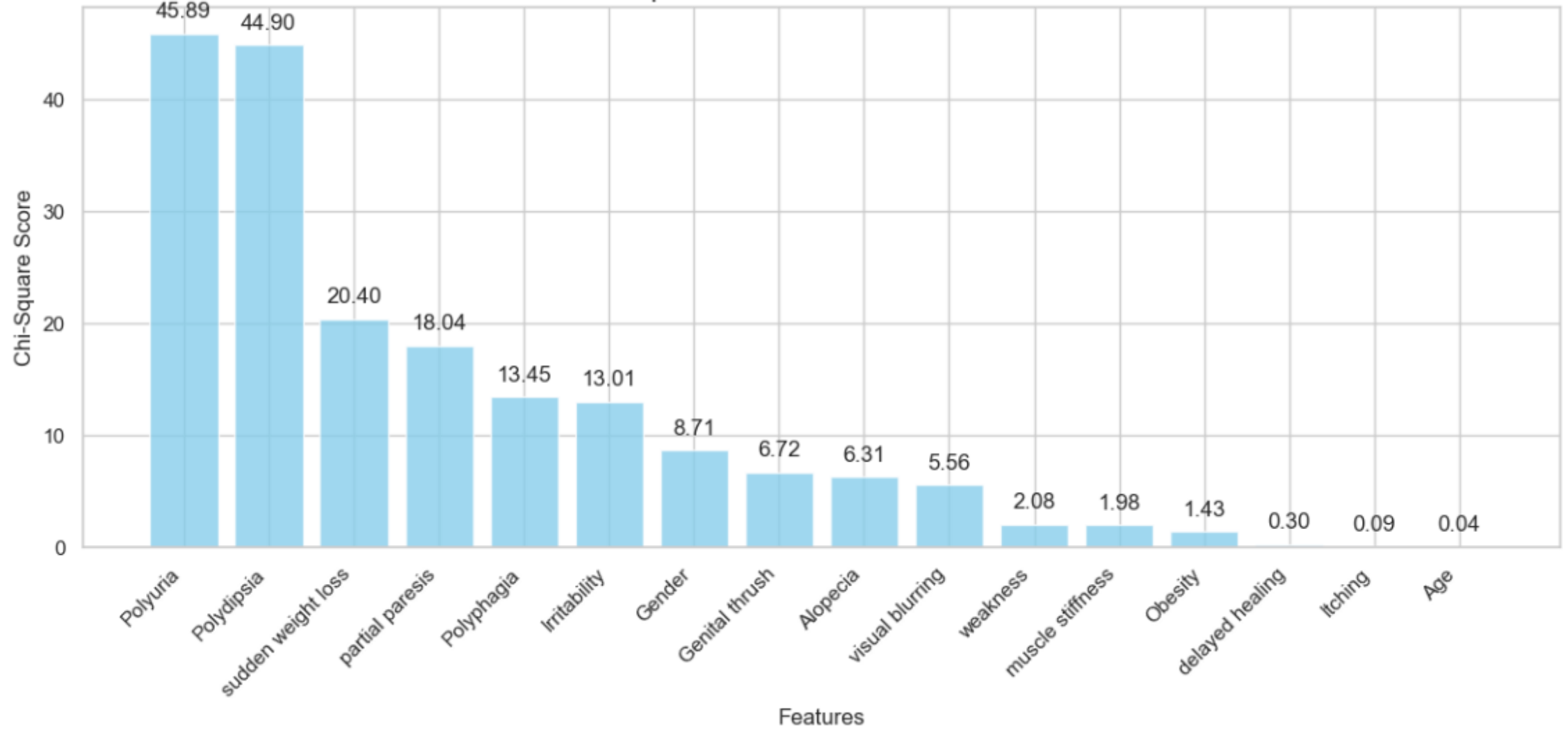


Correlation Matrix

Feature selection:

- Feature selection is a process to identify and retain the most relevant variables for model building while removing redundant or irrelevant features. Key steps in this project included:
- **Chi-Square Test:** Used to evaluate the statistical significance of categorical features with the target variable. Features like **Polyuria** (score: 45.89) and **Polydipsia** (score: 44.90) were identified as top predictors.
- **Impact Categorization:**
 - **High-Impact Features:** Polyuria, Polydipsia, Sudden Weight Loss.
 - **Moderate Features:** Polyphagia, Irritability, Genital Thrush.
 - **Low-Impact Features:** Muscle Stiffness, Obesity.

Chi-Square Scores for Feature Selection



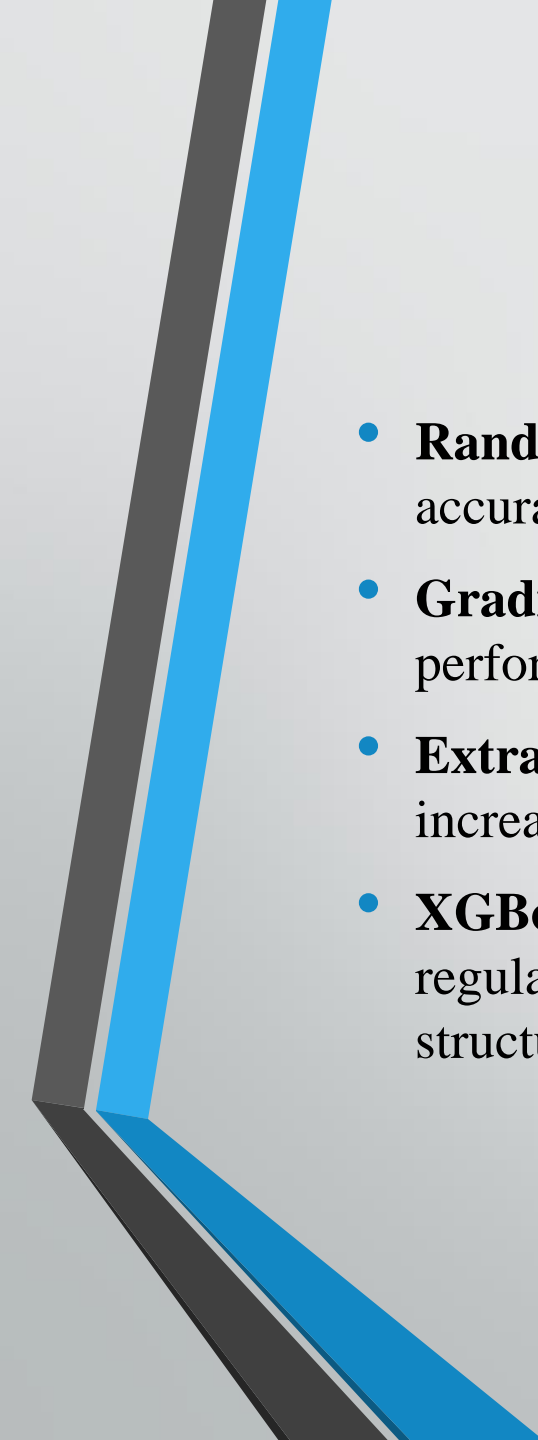
Dimensionality reduction:

- Dimensionality reduction simplifies high-dimensional datasets by reducing the number of features while preserving critical information. In this project:
- **Principal Component Analysis (PCA):**
 - PCA was evaluated to transform the dataset into a lower-dimensional space.
 - Analysis showed 14 out of 16 features retained 95% of the data variance, making dimensionality reduction unnecessary.
- **Impact:**
 - Simplifies data, reduces computational complexity, and minimizes overfitting risks in high-dimensional datasets.
 - In this case, the dataset was already compact, so dimensionality reduction was not required for effective modeling.
- This step ensured the dataset was optimized without unnecessary transformations.



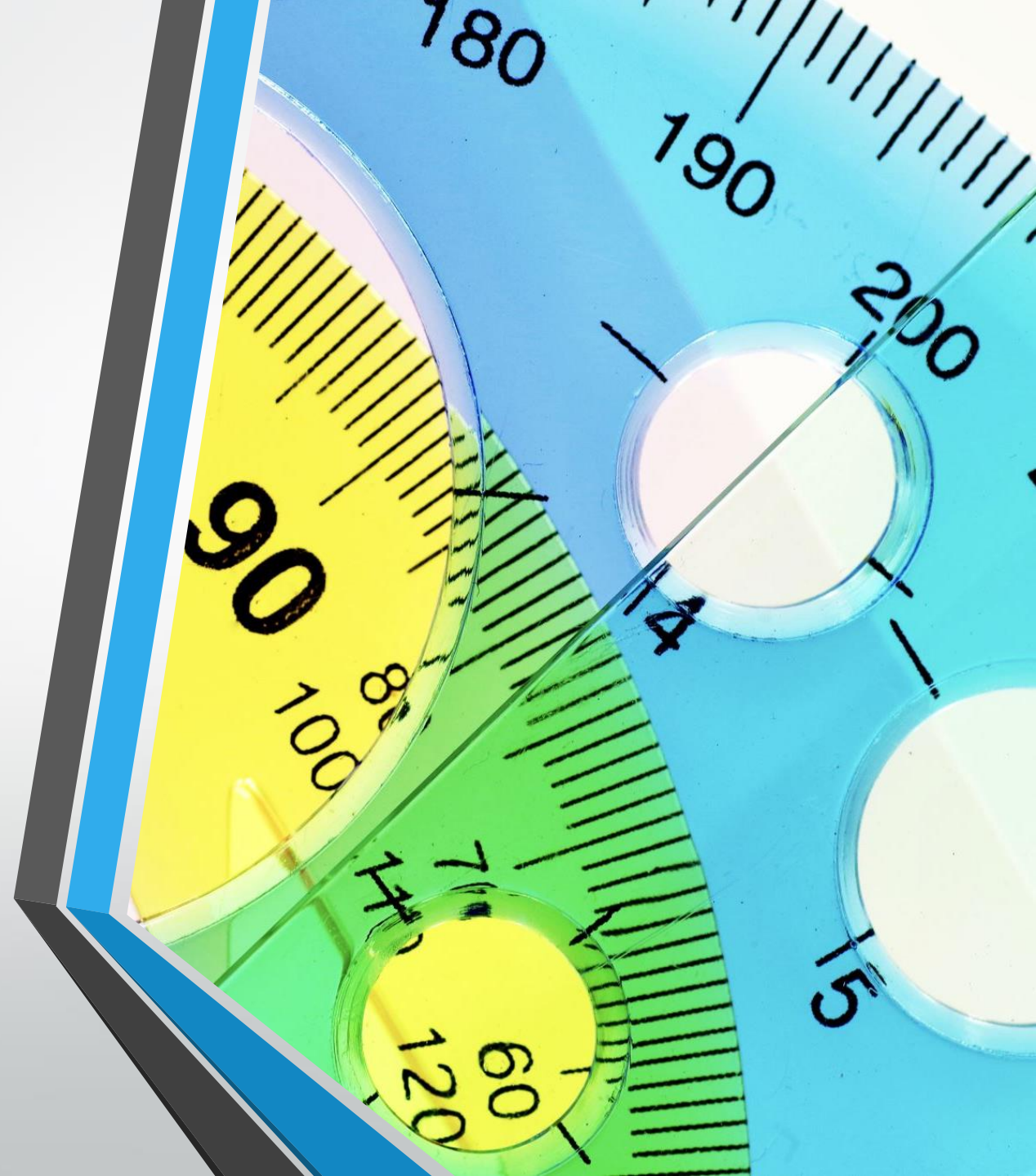
Classification model:

- Classification models were used to predict diabetes status (Diabetic, Pre-Diabetic, or Healthy) based on key features. Various algorithms were implemented and evaluated:
- **Logistic Regression:** A simple, interpretable model that estimates the probability of diabetes using a sigmoid function.
- **Decision Tree:** A tree-based model that splits data into branches for decision-making, offering straightforward interpretation.

- 
- **Random Forest:** An ensemble method combining multiple decision trees to improve accuracy and reduce overfitting.
 - **Gradient Boosting:** Sequentially builds models to minimize errors, ensuring high performance.
 - **Extra Trees:** Similar to Random Forest but uses randomized splits for feature thresholds, increasing diversity and robustness.
 - **XGBoost:** XGBoost enhances performance through optimized tree construction, regularization to prevent overfitting, and support for parallel processing, making it ideal for structured data tasks.

Hyperparameter tuning:

- Hyperparameter tuning is the process of optimizing a model's parameters to achieve the best performance. For this project, key techniques were applied to fine-tune classification models:
- **Methods Used:**
 - **Grid Search:** Tested different combinations of hyperparameters systematically to find the optimal values.
 - **Random Search:** Explored a random subset of hyperparameters for faster tuning with large parameter spaces.



Performance metrics



Accuracy: The ratio of correctly predicted instances to the total instances, indicating overall correctness.



Precision: The ratio of true positive predictions to the total predicted positives, measuring the accuracy of positive predictions.



Recall (Sensitivity): The ratio of true positive predictions to the total actual positives, indicating the model's ability to identify all relevant instances.



F1 Score: The harmonic mean of precision and recall, providing a balance between the two metrics, especially useful for imbalanced dataset.



Area Under Curve: AUC is a single number that summarizes how well a model can distinguish between positive and negative instances.

Evaluation
metrics before
hyperparameter
tuning:

	Model	Accuracy	Precision	Recall	F1 Score	AUC
0	Logistic Regression	0.823529	0.825000	0.942857	0.880000	0.944643
1	Decision Tree	0.882353	0.891892	0.942857	0.916667	0.846429
2	Support Vector Machine	0.901961	0.894737	0.971429	0.931507	0.971429
3	Random Forest	0.921569	0.918919	0.971429	0.944444	0.975893
4	Gradient Boosting	0.901961	0.894737	0.971429	0.931507	0.951786
5	Extra Trees	0.921569	0.918919	0.971429	0.944444	0.989286
6	XGBoost	0.882353	0.891892	0.942857	0.916667	0.951786

Evaluation
metrics before
hyperparameter
tuning:

	Model	Accuracy	Precision	Recall	F1-Score	AUC
0	Logistic regression	0.862745	0.850000	0.971429	0.906667	0.950000
1	Decision Tree	0.882353	0.891892	0.942857	0.916667	0.950000
2	SVM	0.882353	0.871795	0.971429	0.918919	0.967857
3	Random forest	0.901961	0.916667	0.942857	0.929577	0.977679
4	Gradient Boosting	0.862745	0.888889	0.914286	0.901408	0.950000
5	Extra trees	0.921569	0.918919	0.971429	0.944444	0.983929
6	XGBoost	0.901961	0.894737	0.971429	0.931507	0.951786



Model selection:

- Extra Trees stand out as the most promising models due to their strong performance across all metrics. These model would likely be the preferred choices for a classification problem. So, I would like to choose Extra trees model.
- Here are the reasons to choose the Extra Trees model:
 1. Superior performance across metrics.
 2. High precision and recall.
 3. Balanced model complexity.
 4. Robustness to overfitting.
 5. Scalability



A decorative graphic on the left side of the slide. It features two wooden pawns, one white and one red, standing on a white surface. A blue and black diagonal line runs from the bottom left towards the top right, separating the pawns from the text area.

Conclusion

This project successfully developed a predictive model for diabetes risk estimation, leveraging AI and machine learning techniques. The model demonstrated high accuracy in predicting the likelihood of diabetes, offering valuable insights into the lifestyle and healthcare factors that influence diabetes risk.

Future recommendations and extensions:

Future recommendations include expanding the dataset to include additional lifestyle and environmental factors, exploring advanced machine learning models for improved accuracy, and integrating real-time health data for continuous risk assessment. Further research could also focus on testing the model's generalizability across diverse populations and healthcare settings.





THANK YOU