Pima Indian Heritage (Diabetes) Data Analysis

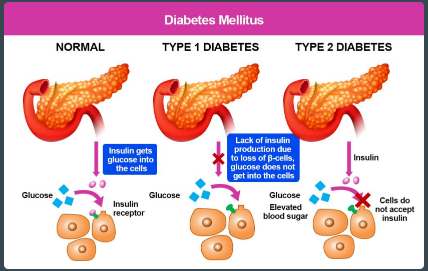
Group 7 - Project 4

Yang, Edward - Rincon, Victor - Locke, Khalil - Cetin, Ugur

Table of Contents

- Introduction
- What is our data set?
- Data Cleaning
- Machine Learning
- Demo/Tableau
- Conclusion
- Bias and Limitations
- Call to Action
- Future Work
- Q & A

Introduction


Diabetes Mellitus
NORMAL
TYPE 1 DIABETES
TYPE 2 DIABETES
Insulin gets glucose into the cells
Glucose
Insulin receptor
Lack of insulin production due to loss of β-cells, glucose does not get into the cells
Glucose
Glucose
Elevated blood sugar
Insulin
Cells do not accept insulin

- What is Diabetes?
    - Diabetes is a chronic health condition characterized by high blood sugar (glucose) levels due to the body's inability to properly produce or use insulin. Insulin is a hormone produced by the pancreas that regulates blood sugar by helping glucose enter cells for energy.
    - Types of Diabetes
        - Type 1
            - Autoimmune, insulin producing cells are attacked, requires lifelong insulin therapy
        - Type 2
            - Most common type, Results from insulin resistance, where the body's cells do not respond effectively to insulin, often combined with reduced insulin production, Managed through lifestyle changes, oral medications, and sometimes insulin
        - Gestational
            - Occurs during pregnancy in women without a prior history of diabetes.
            - Typically resolves after childbirth but increases the risk of developing Type 2 diabetes later in life.

- Purpose of this project was to use machine learning where we take values for specific features to determine if someone would have diabetes.
    - Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age were the featured values.

Data

Kaibab - Paiute Tribe
Havasupai Tribe
Navajo Nation
COCONINO
Hopi Tribe
Hualapai Tribe
MOJAVE
NAVAJO
APACHE
Fort Mojave Tribe
YAVAPAI
Yavapai-Apache Nation
Yavapai-Prescott Indian Tribe
Zuni Pueblo
Tonto Apache Tribe
Colorado River Indian Tribe
GILA
White Mountain Apache Tribe
LA PAZ
Fort McDowell Yavapai Nation
MARICOPA
Salt River Pima-Maricopa Indian Community
San Carlos Apache Tribe
GREENLEE
Gila River Indian Community
Ak Chin Indian Community
YUMA
PINAL
Quechan Tribe
GRAHAM
Cocopah Tribe
Tohono O'odham Nation
Pascua Yaqui Tribe
PIMA
COCHISE
SANTA CRUZ

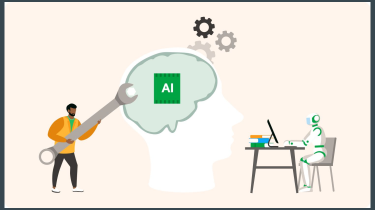
- The dataset we used was called Pima Indians Diabetes.
  - Reasoning
    - High prevalence of type 2 diabetes among the Pima Indians.
    - Rates are among the highest of any population in the world.
    - Linked to genetic predisposition and lifestyle changes.
      - Genetic predisposition related to insulin resistance.
      - Sedentary lifestyle replacing traditional active farming life.
        - Traditionally relied on farming: corn, beans, squash, and later wheat.
        - Transition from traditional diets to Western-style diets.

Data Cleaning

- No null values
- No categorical, only numerical values
- Lot of zeroes, filled with the median due to:
  - Median is not affected by extreme values (outliers)
  - Values replaced by the median ensures the substitution reflects the dataset's central trend
  - The mean, in contrast, could result in a distorted dataset because zeros would disproportionately pull the mean downward
  - In datasets with many zeroes or outliers, the mean can be difficult to interpret because it might not represent any actual data point well.

Machine Learning

- Prior to pipeline, SMOTE (Synthetic Minority Over-sampling Technique) was applied.
  - Used this technique solely dataset only consisted of females, so this would help generate “males” into our dataset as well increasing the size of overall sample
  - Works by generating synthetic samples for the minority class rather than just oversampling the existing data points.
  - Creates these new instances by taking the nearest neighbors of an existing data point and generating synthetic samples between them
  - Helps balance the class distribution, which can improve the performance of machine learning models.
- Logistic Regression as our pipeline
  - Binary classification (1/0, yes/no)
  - Highly interpretable, meaning you can understand how each feature influences the prediction.
  - Slight overfitting AUC .85 vs .81
  - Accuracy of 77%

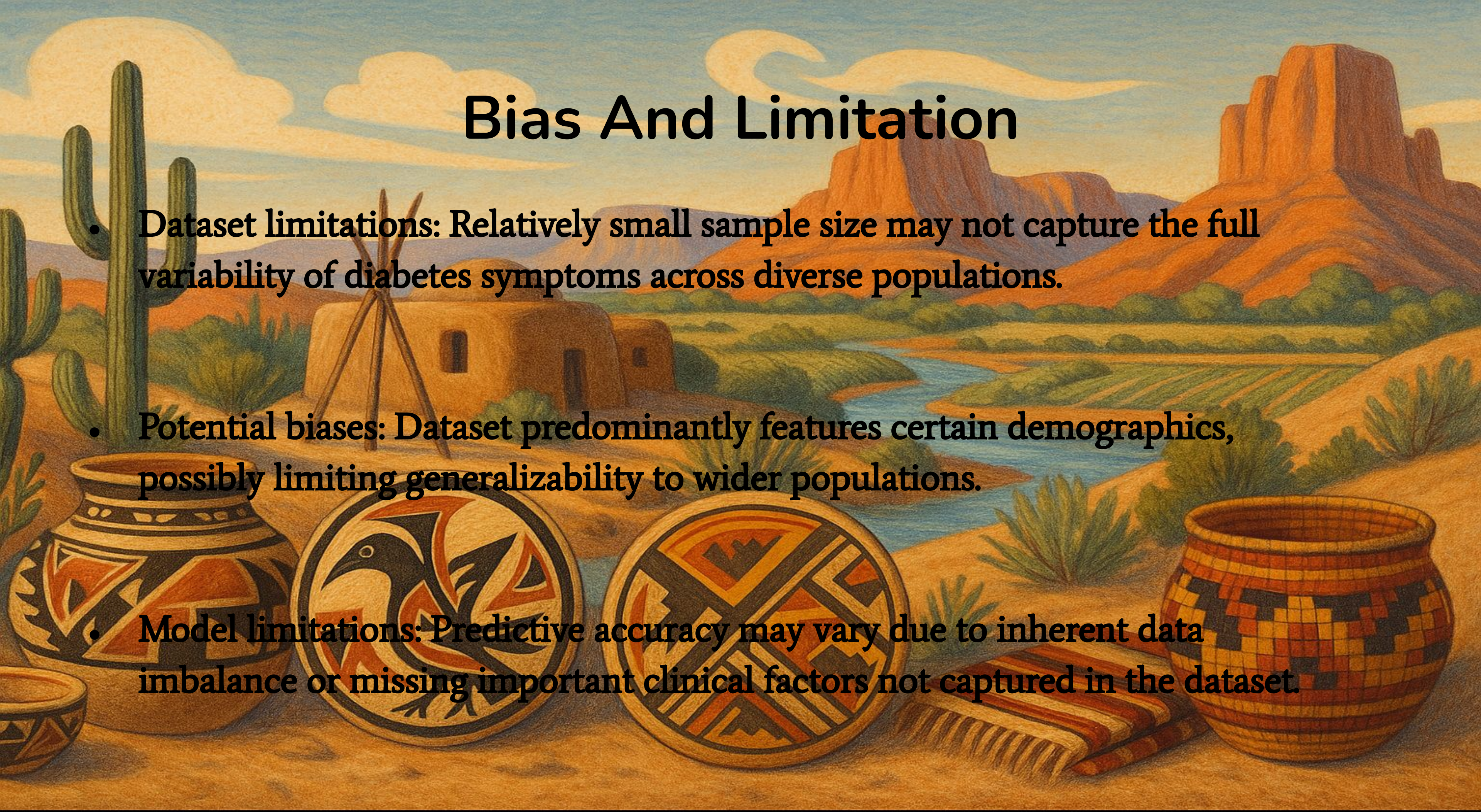


Conclusion

- The analysis of diabetes data reveals significant relationships between diagnostic measurements and diabetes occurrence.

- Predictive modeling successfully demonstrated potential diagnostic accuracy, highlighting the importance of careful feature selection.

- Data-driven analysis provides valuable insights and can substantially improve clinical decision-making and patient outcomes.



Bias And Limitation

- Dataset limitations: Relatively small sample size may not capture the full variability of diabetes symptoms across diverse populations.

- Potential biases: Dataset predominantly features certain demographics, possibly limiting generalizability to wider populations.

- Model limitations: Predictive accuracy may vary due to inherent data imbalance or missing important clinical factors not captured in the dataset.

Call to Action

- Encourage healthcare providers to integrate data-driven predictive tools into routine diagnostic procedures.
- Advocate for collecting comprehensive and diverse datasets to enhance model accuracy and applicability.
- Promote training for medical professionals on interpreting and applying predictive analytics in clinical environments.

Future Work

- Expand the dataset to include diverse demographics and additional clinical parameters for a comprehensive analysis.

- Explore advanced machine learning and deep learning algorithms to potentially improve predictive accuracy.

- Collaborate with healthcare professionals to validate models in clinical trials or real-world clinical settings.

Q&A