Project presentation

Paula Lozano Gonzalo

Predict diabetes using patient clinical data.

Dataset

SOURCE

KEY FEATURES

PREPROCESSING

Kaggle dataset

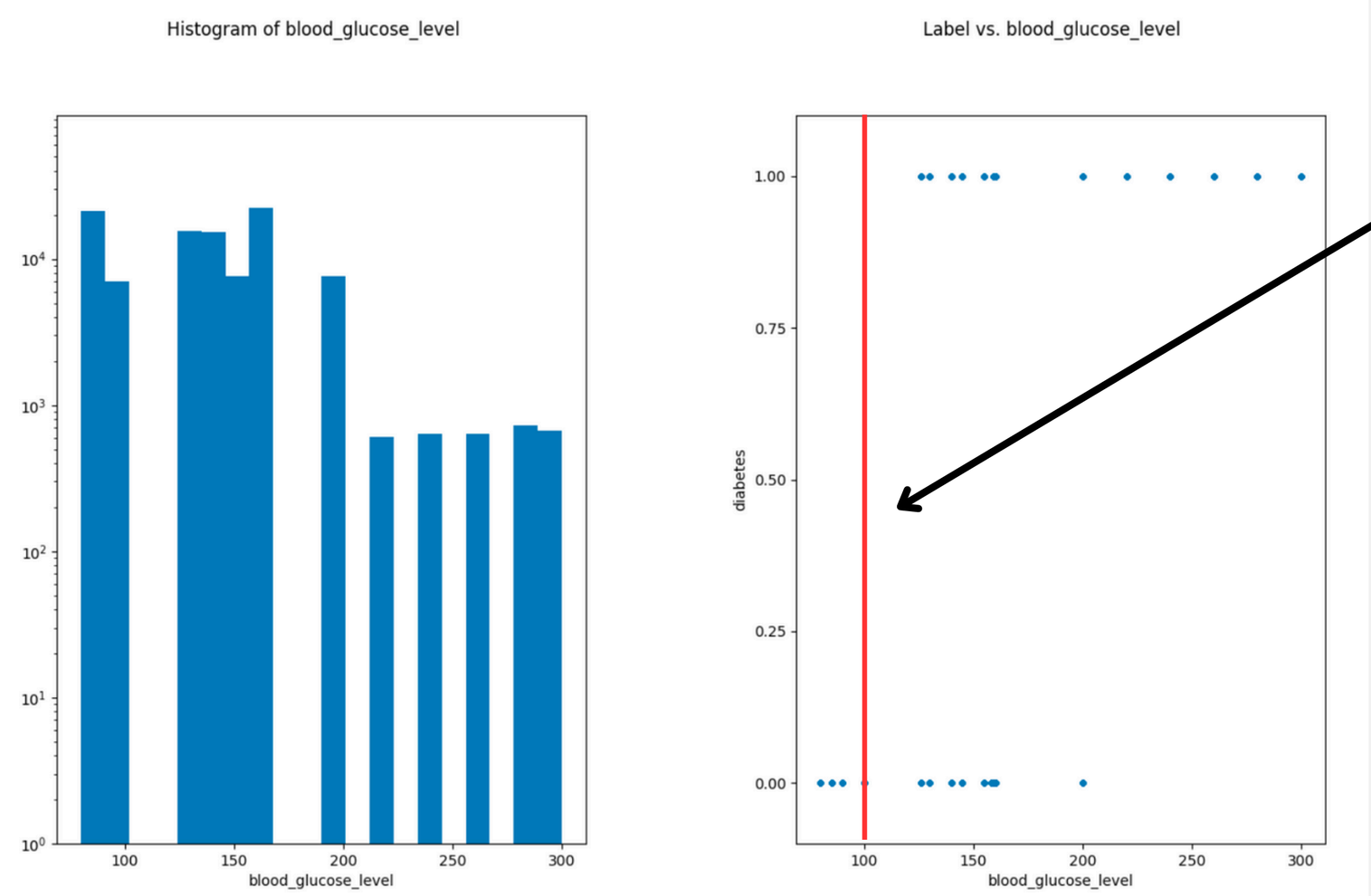
Features

1. Year
2. Age
3. Gender
4. Location
5. Race
6. Hypertension
7. Heart Disease
8. Smoking History
9. BMI
10. HbA1c Levels
11. Blood Glucose Levels
12. Clinical Notes

Key Features

Blood Glucose Level
&
HbA1c

Blood Glucose Level



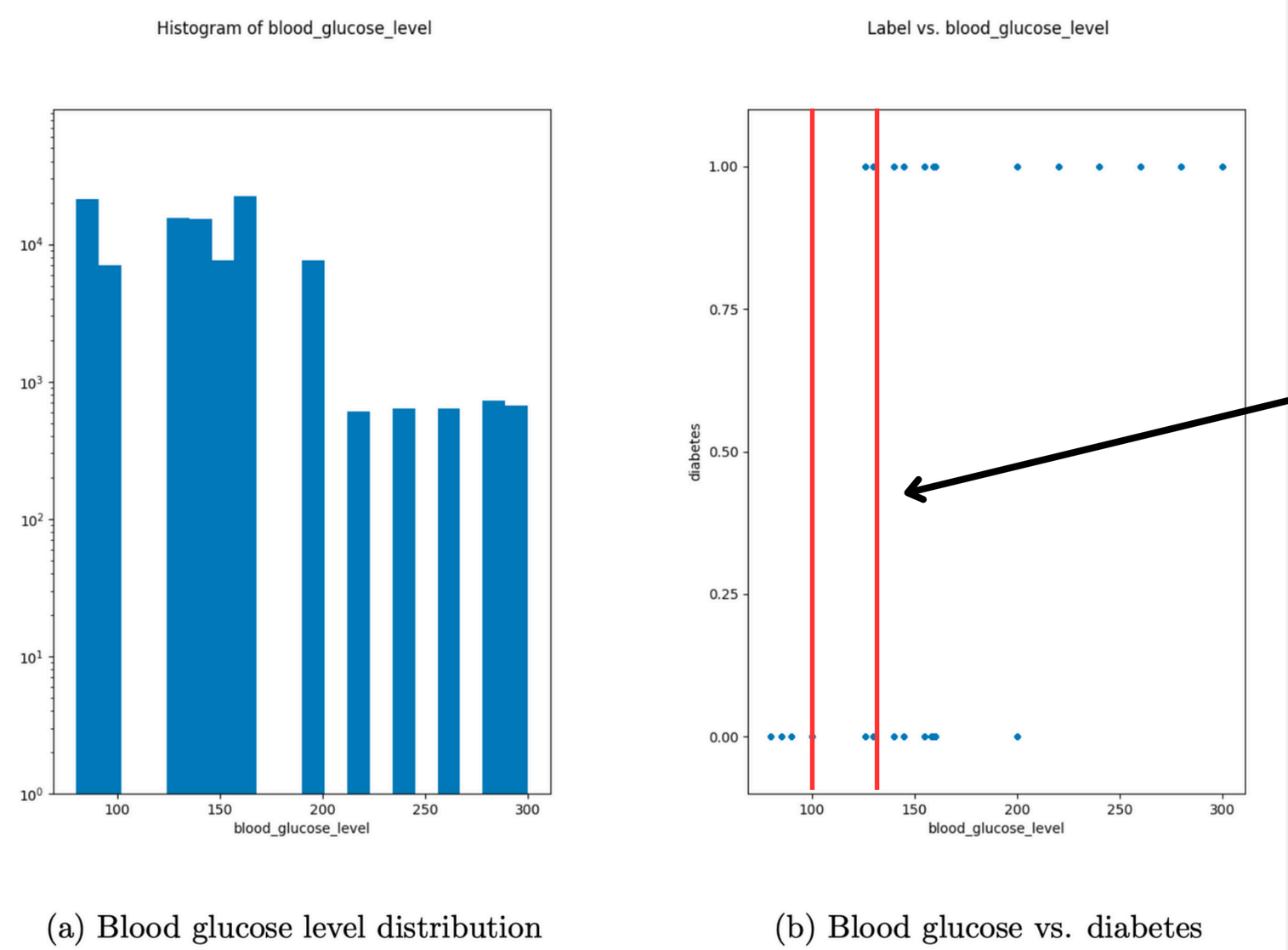
(a) Blood glucose level distribution

(b) Blood glucose vs. diabetes

NORMAL:

<100 mg/dL

Blood Glucose Level



(a) Blood glucose level distribution

(b) Blood glucose vs. diabetes

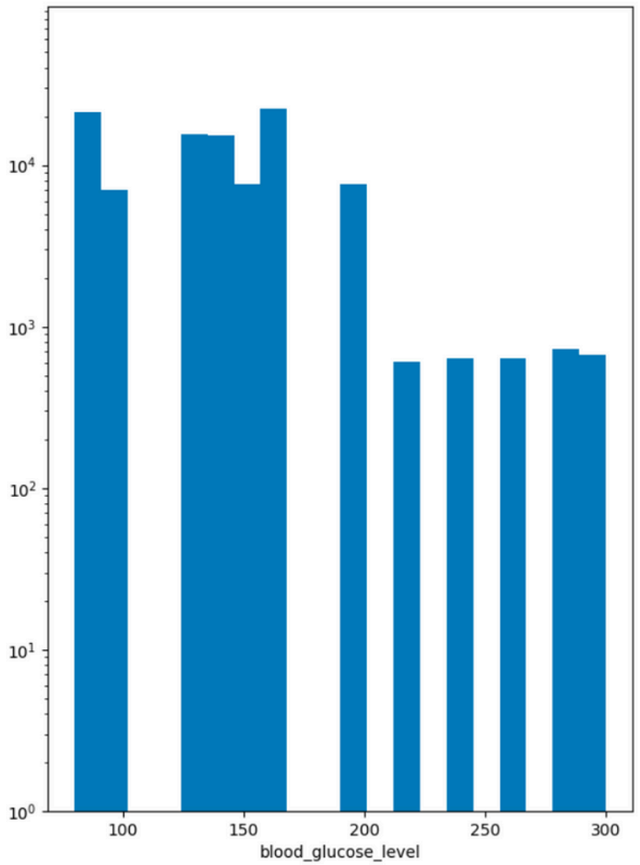
NORMAL:

<100 mg/dL

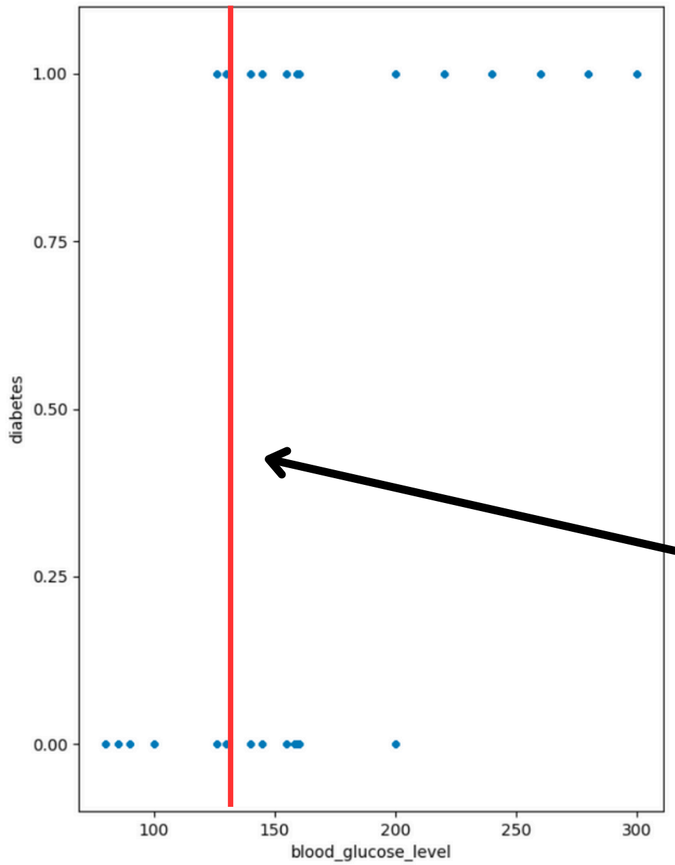
PREDIABETES:

100–125 mg/dL

Blood Glucose Level



(a) Blood glucose level distribution



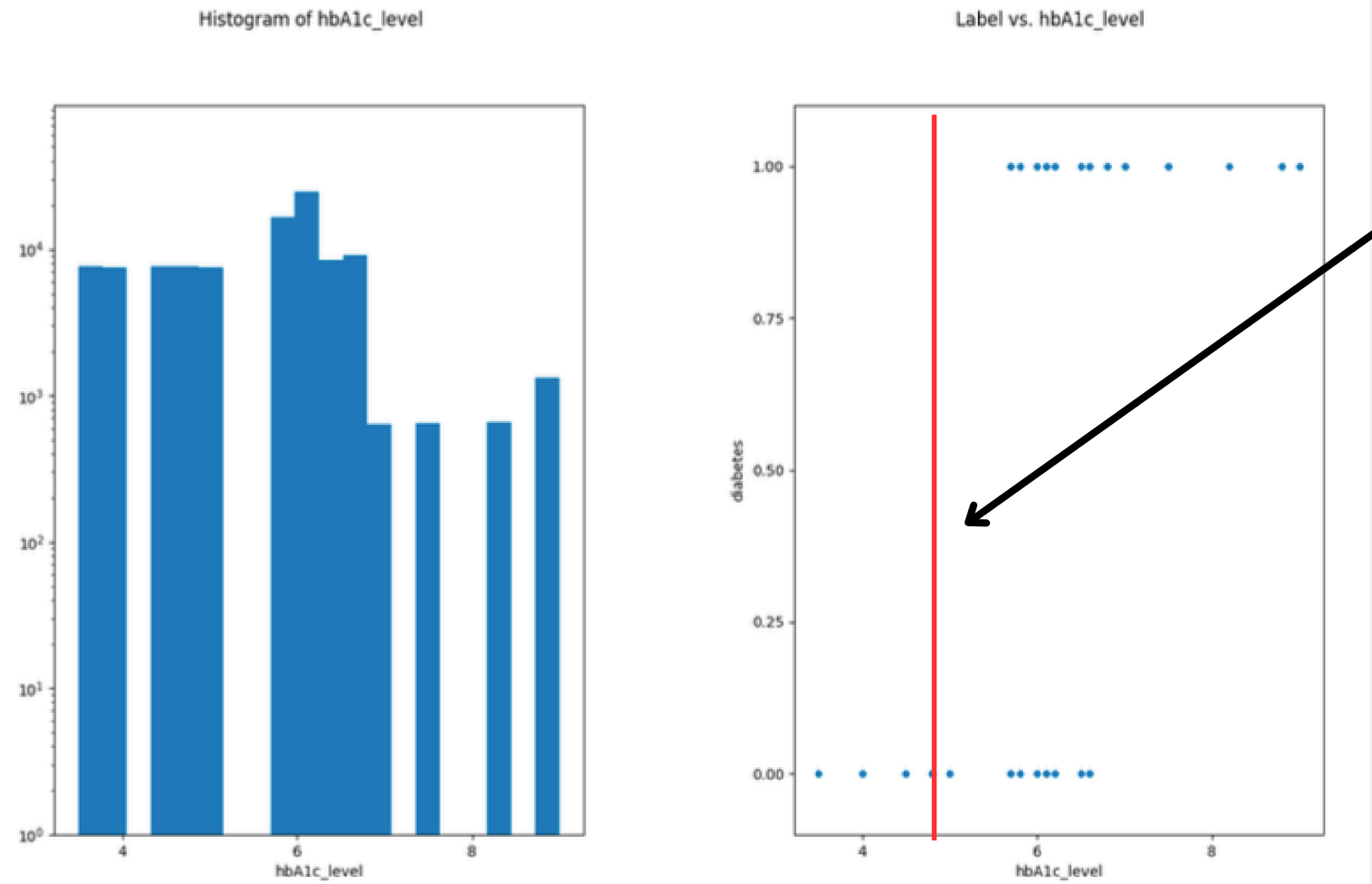
(b) Blood glucose vs. diabetes

NORMAL:
<100 mg/dL

PREDIABETES:
100–125 mg/dL

DIABETES:
>126 mg/dL

HbA1c (Average blood glucose levels over approximately 3 months)

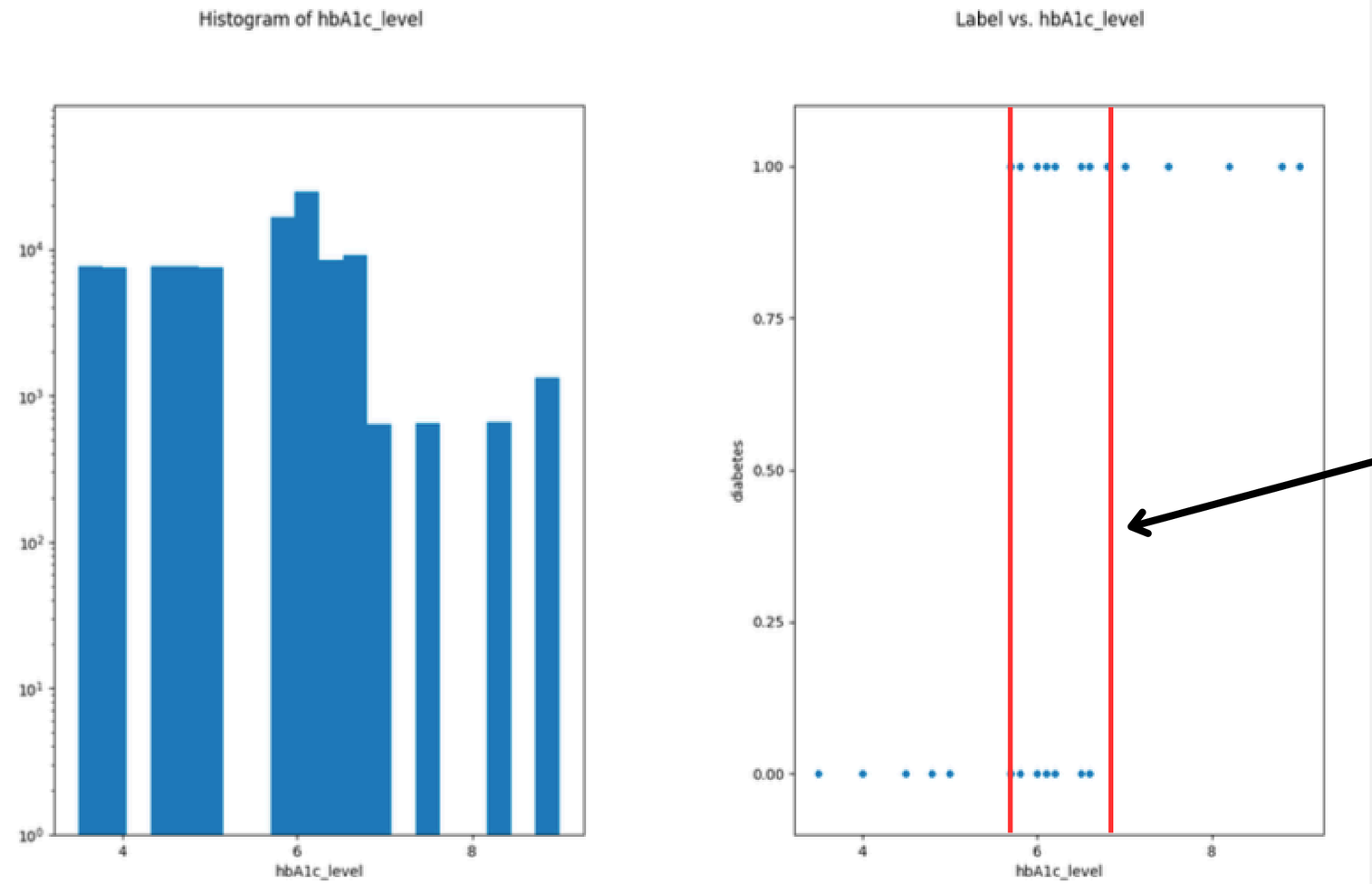


(a) HbA1c level distribution

(b) HbA1c vs. diabetes

NORMAL: <5.7%

HbA1c (Average blood glucose levels over approximately 3 months)



(a) HbA1c level distribution (b) HbA1c vs. diabetes

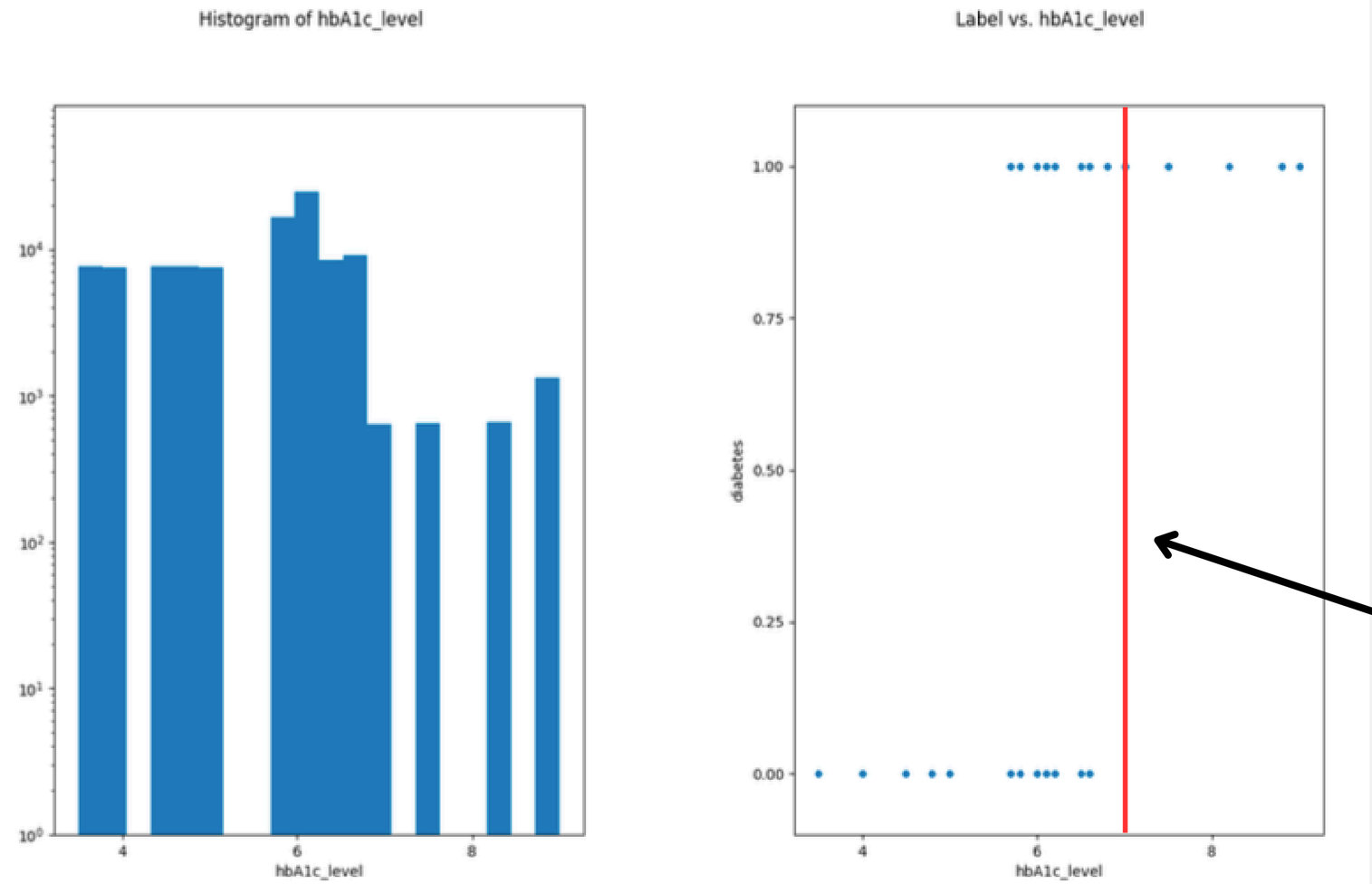
NORMAL:

<5.7%

PREDIABETES:

5.7% - 6.4%

HbA1c (Average blood glucose levels over approximately 3 months)



(a) HbA1c level distribution (b) HbA1c vs. diabetes

NORMAL:
<5.7%

PREDIABETES:
5.7% – 6.4%

DIABETES:
>6.5%

Preprocessing

1. Handled missing values.
2. One hot-encoded categorical features.
3. Added polynomial features.
4. Dropped “clinical notes” feature.

Models Trained

First Models

SGDClassifier

===== Training Data =====

	Predicted No (F)	Predicted Yes (T)
Actual No	64,608	1,307
Actual Yes.	2,232	3,853

Metric	Value
Precision	0.747
Recall	0.633
F1 Score	0.685

Forest Classifier

===== Training Data =====

	Predicted No (F)	Predicted Yes (T)
Actual No	65,887	28
Actual Yes.	1,990	4,095

Metric	Value
Precision	0.993
Recall	0.673
F1 Score	0.802

First Models

SGDClassifier

===== Training Data =====

	Predicted No (F)	Predicted Yes (T)
Actual No	64,608	1,307
Actual Yes.	2,232	3,853

Metric	Value
Precision	0.747
Recall	0.633
F1 Score	0.685

Forest Classifier

===== Training Data =====

	Predicted No (F)	Predicted Yes (T)
Actual No	65,887	28
Actual Yes.	1,990	4,095

Metric	Value
Precision	0.993
Recall	0.673
F1 Score	0.802

Unbalanced Data

Total cases: 100,000
Positive cases: 8,500
Negative cases: 91,500

→

Total cases: 17,000
Positive cases: 8,500
Negative cases: 8,500

Best Models

Forest Classifier

===== Test Data =====

	Predicted No (F)	Predicted. Yes (T)
Actual No	1,544	158
Actual Yes	153	1,545

Metric	Value
Precision	0.907
Recall	0.910
F1 Score	0.909

AdaBoost

===== Test Data =====

	Predicted No (F)	Predicted Yes (T)
Actual No	1,511	191
Actual Yes.	149	1,549

Metric	Value
Precision	0.890
Recall	0.912
F1 Score	0.901

Neural Network

===== Test Data =====

	Predicted No (F)	Predicted Yes (T)
Actual No	1,210	492
Actual Yes	51	1,647

Metric	Value
Precision	0.770
Recall	0.970
F1 Score	0.858

Next Steps

01

Generating synthetic data instead of dropping so much data.

02

Try with more models and different hyperparameters.

Thank you very much!