CIND 820 – CAPSTONE PROJECT

PREDICTION OF DIABETES USING ML TECHNIQUES

Muhammad Maqsud Hussain
m5hussain@torontomu.ca
Student No. 500978290

Supervised By:
Dr. Ceni Babaoglu

INTRODUCTION

DIABETES

- Chronic disease
- 1 in every 10 adults in the world is living with diabetes.
- 30% of Canadians live with diabetes or prediabetes (Diabetes Canada).
- Risks:
  - Cardiovascular diseases- 3 times
  - End-stage renal diseases-12 times
  - Non-traumatic lower limb amputation- 20 times

Why Early Prediction?

Prevent permanent damage to the heart, kidneys, eyes, nerves, blood vessels, and other vital organs.

How?

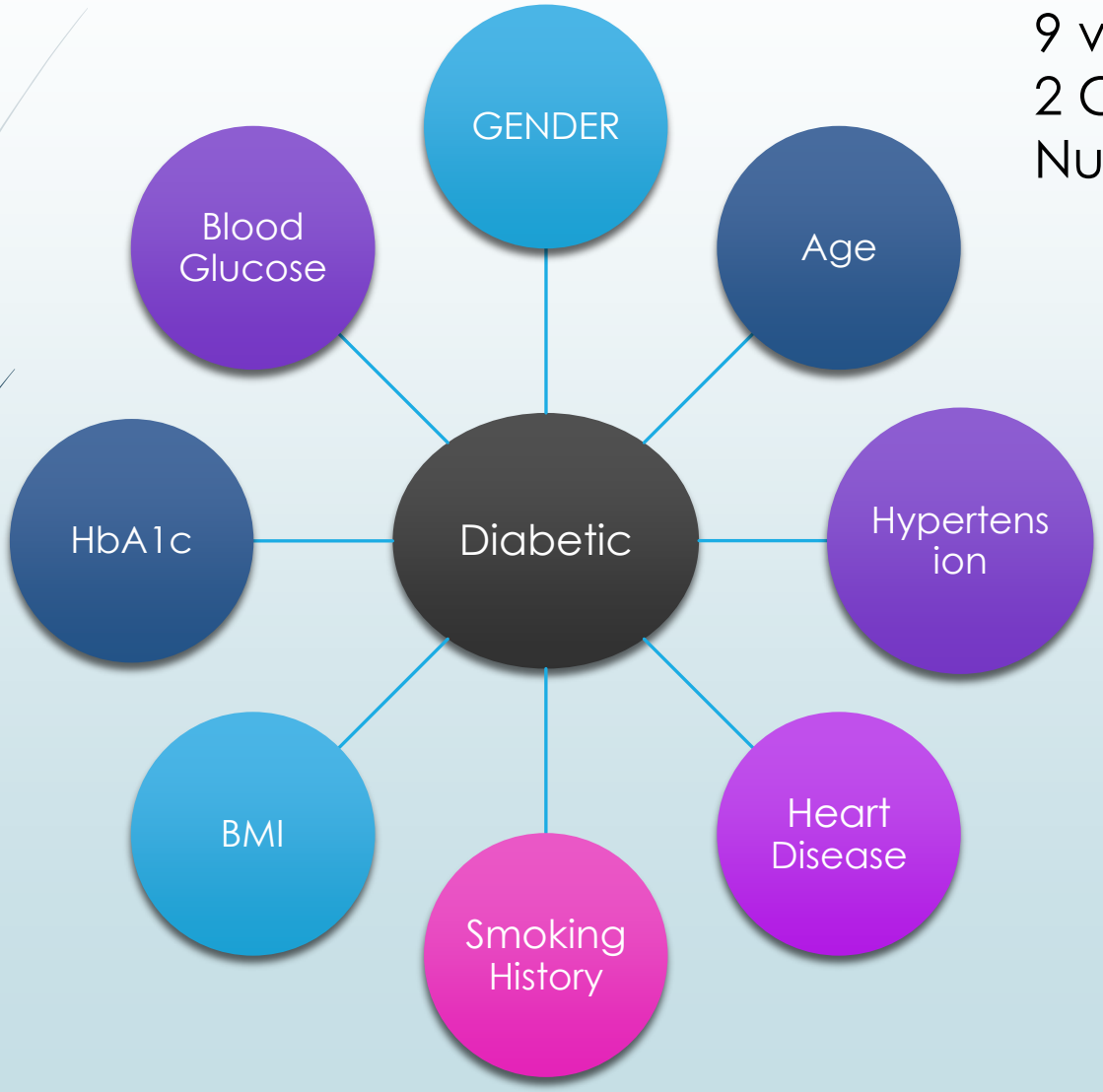
Machine Learning

RESEARCH QUESTIONS

- Which factors are important?
- Is blood sugar level of an obese/overweight person always higher than a person with normal BMI?
- How accurately can diabetes be predicted using ML?
- Which model works best?

THE DATASET

100,000 records
9 variables
2 Categorical, 7 Numerical



METHODOLOGY

Data Preprocessing

- Initial Analysis
- Univariate Analysis
- Bivariate Analysis

Model Development

- Logistic regression Classifier
- K-nearest neighbor regression
- Decision tree classifier
- Random forest classifier
- Support vector modeling

Application and Modification

- Initial Iteration
- Cross-Validation

Result Analysis

- Evaluation of Models
- Comparison of ML models
- Limitations
- Recommendations

DATA PREPROCESSING

INITIAL ANALYSIS

- Missing Values

  No Missing Values

- Duplicate Records
  - 3854 records among the 100,000 rows
  - Not removed as records are anonymous

- Statistical Measures

Count, mean, std etc. was measured and analyzed.

	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes
count	100000.000000	100000.00000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000
mean	41.885856	0.07485	0.039420	27.320767	5.527507	138.058060	0.085000
std	22.516840	0.26315	0.194593	6.636783	1.070672	40.708136	0.278883
min	0.080000	0.00000	0.000000	10.010000	3.500000	80.000000	0.000000
25%	24.000000	0.00000	0.000000	23.630000	4.800000	100.000000	0.000000
50%	43.000000	0.00000	0.000000	27.320000	5.800000	140.000000	0.000000
75%	60.000000	0.00000	0.000000	29.580000	6.200000	159.000000	0.000000
max	80.000000	1.00000	1.000000	95.690000	9.000000	300.000000	1.000000

DATA PREPROCESSING

UNIVARIATE ANALYSIS

a)Gender:

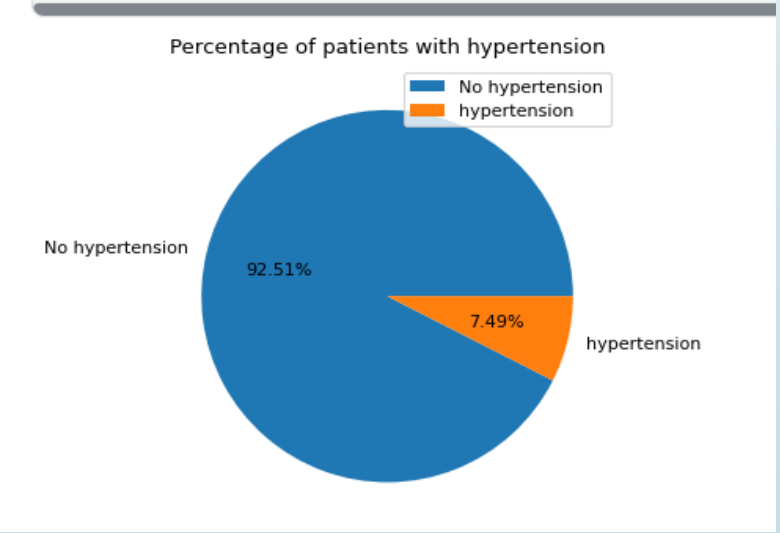
- **Male-41,430, Female-58,552**
- Other-18 (REMOVED)

b) AGE:

- **Range:** Newborn to 80 years
- **Mean-**41.88 years

c) Hypertension:

7.49% have hypertension



DATA PREPROCESSING

UNIVARIATE ANALYSIS Contd.

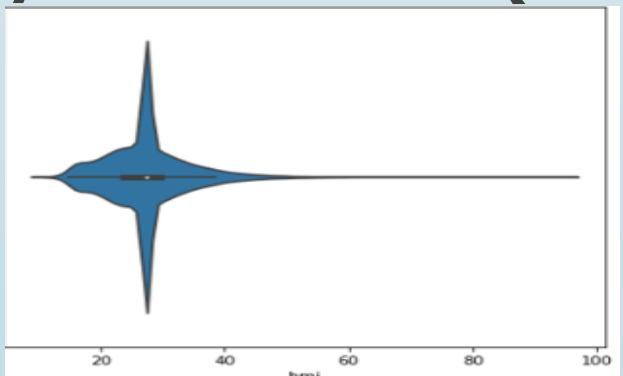
d) Heart Disease:

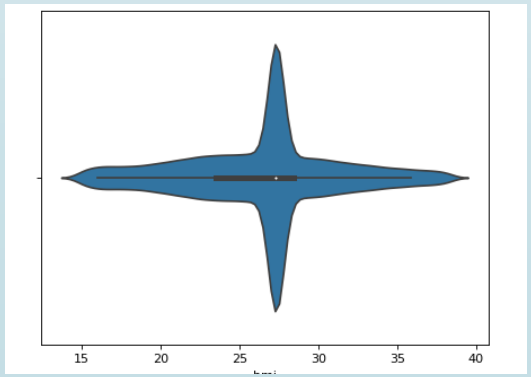
Heart disease -3,942 (3.94%)

e) Smoking History:

smoking_history	
No Info	35810
never	35092
former	9352
current	9286
not current	6439
ever	4003

f) Body Mass Index (BMI):





DATA PREPROCESSING

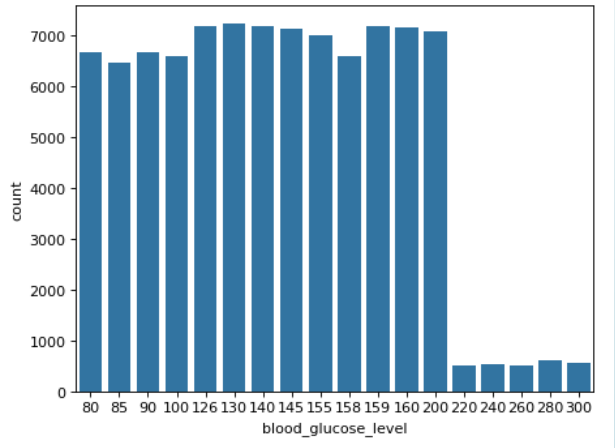
UNIVARIATE ANALYSIS Contd.

g) HbA1c Level:

- Range: **3.5% to 9%**
- **Below 5.*7%*** *non-diabetic*, ***5.7%-6.4%*** *prediabetic* and **6.5% or above-***diabetic*

e) Blood Glucose Level:

- Range: **80 mg/dl to 300 mg/dl**



f) Diabetes(Target Feature):

- Diabetes in this dataset is **7.56%**
- Global percentage of diabetic patient is **9.30%**

DATA PREPROCESSING

BIVARIATE ANALYSIS:

HbA1c_level vs. diabetes:

HbA1c level	Initial diagnosis	Actual diagnosis
< 5.7	Normal	100% no diabetes
5.7 – 6.4	Prediabetes	7.47% have diabetes
>= 6.5	Diabetes	23.64% have diabetes

BMI vs diabetes:

BMI	Category	Prediction
=< 18.5	Underweight	0.7% have diabetes
18.5 – 24.9	Normal	3.85% have diabetes
25 – 29.9	Overweight	7.88% have diabetes
>= 30	Obesity	15.94% have diabetes

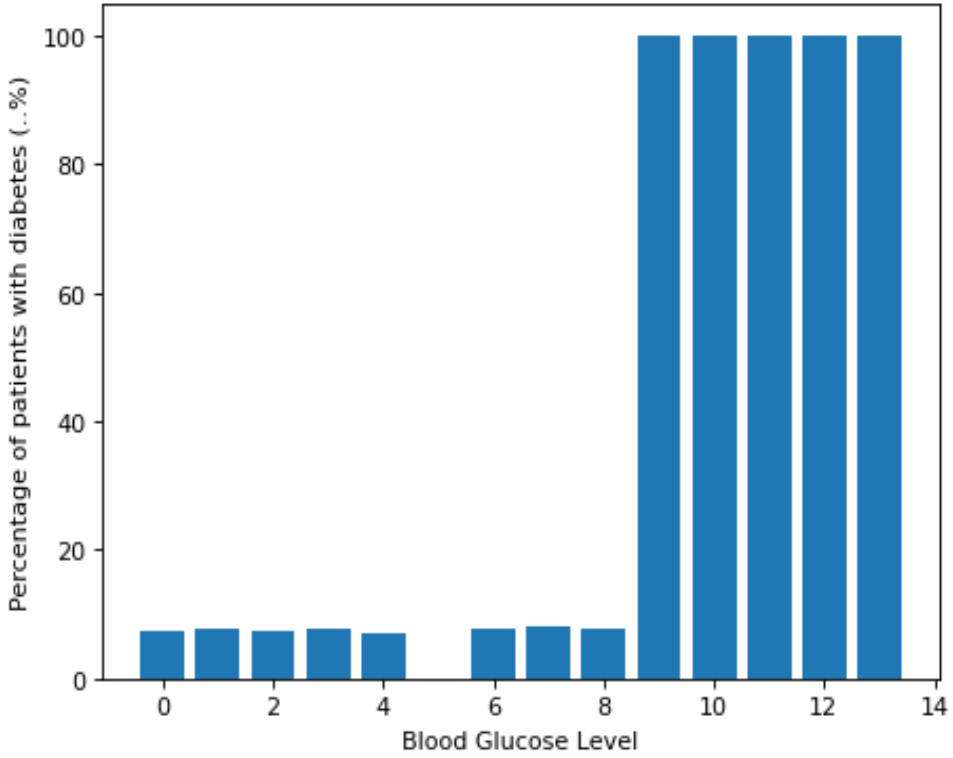
DATA PREPROCESSING

BIVARIATE ANALYSIS Contd.

Blood Glucose Level vs diabetes:

	blood_glucose_level	diabetes	total	percentage
0	126	527	7190	7.33
1	130	566	7231	7.83
2	140	522	7178	7.27
3	145	543	7142	7.60
4	155	482	7019	6.87
5	158	0	6599	0.00
6	159	544	7197	7.56
7	160	584	7150	8.17
8	200	542	7081	7.65
9	220	500	500	100.00
10	240	537	537	100.00
11	260	518	518	100.00
12	280	601	601	100.00
13	300	556	556	100.00

Blood Glucose Level	Category
=< 99	Normal
100 – 125	Prediabetes
>= 126	Diabetes



MODELS

- Logistic Regression Classifier
- K-nearest neighbor regression
- Decision tree classifier
- Random forest classifier
- Support vector modeling

- **Test-Train Split :**

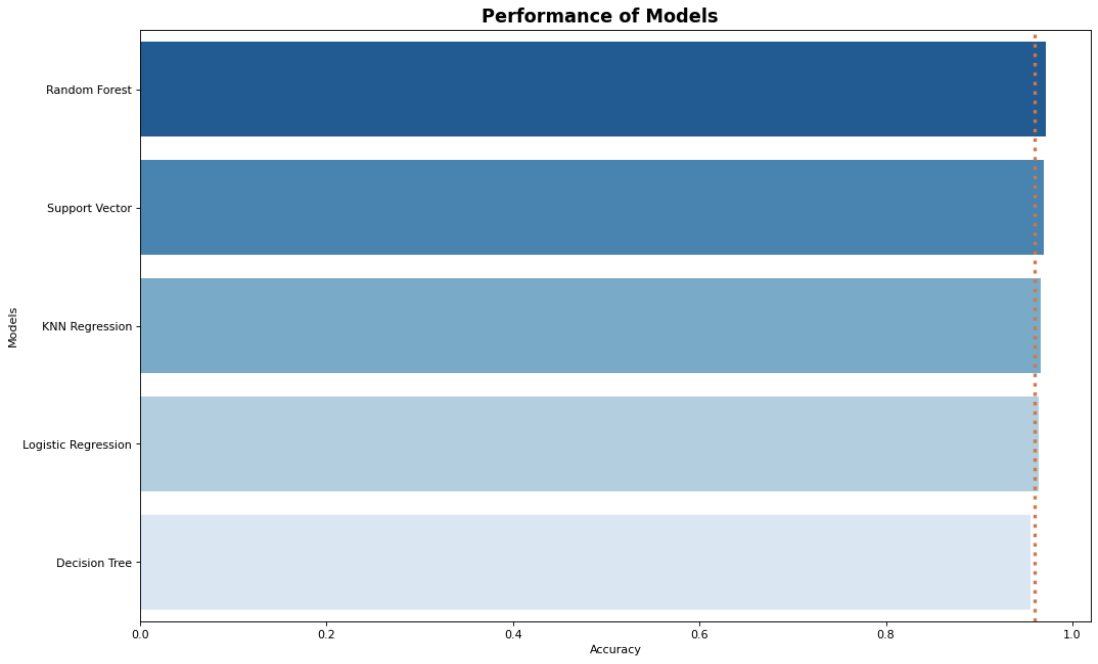
  Ratio: 20-80

- **Standardization and Transformation:**

  using StandardScaler() function

RESULTS-ITERATION

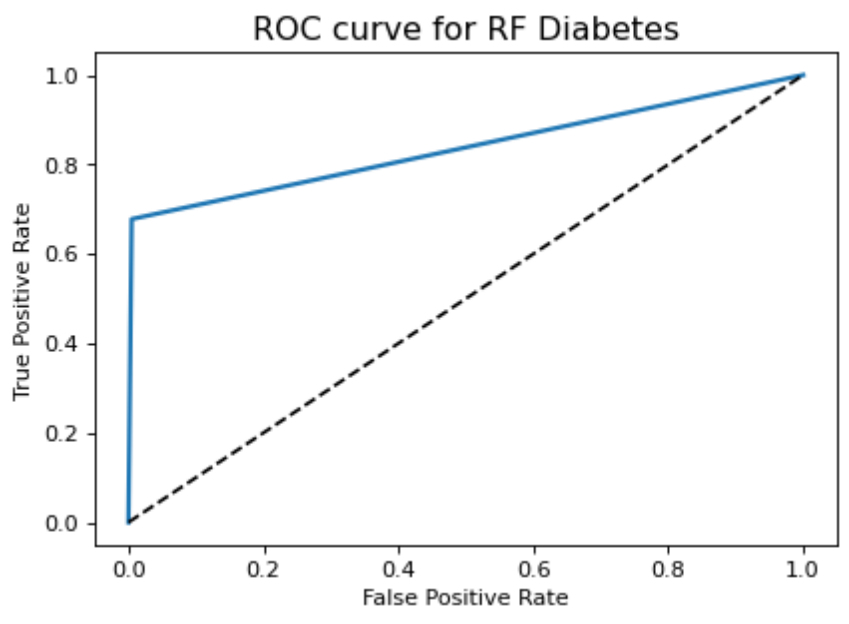
	Model	Accuracy	Precision	Recall	F1 Score
3	Random Forest	0.971367	0.920311	0.677881	0.780709
4	Support Vector	0.969537	0.985965	0.603436	0.748668
1	KNN Regression	0.965554	0.896335	0.612742	0.727891
0	Logistic Regression	0.963455	0.864837	0.609162	0.714826
2	Decision Tree	0.955436	0.695533	0.724409	0.709677



RESULTS- CROSS VALIDATION

- Cross-validation using 10-folds

Algorithm	Mean Accuracy Score	Standard Deviation
Random Forest	97.16 %	0.18%
KNN	96.05%	0.21 %
Decision Tree	95.73%	0.29%
SVM	95.14%	0.34%



ROC AUC- 0.8366

CV ROC AUC- 0.9566

CONCLUSION

ANSWERS TO THE RESEARCH QUESTIONS

- ✓ DATA PREPARATION- IMPORTANT STEP
- ✓ BEST MODEL- RANDOM FOREST
- ✓ CROSS VALIDATION – IMPROVED PERFORMANCE
- ✓ MOST IMPORTANT FACTOR- HbA1c level
- ✓ BMI- Some impact on being a diabetic

LIMITATIONS & RECOMMENDATIONS

- ✓ Dataset lacks important demographic, genetical information
- ✓ Test the model with other datasets
- ✓ Artificial Neural Network (ANN) models

QUESTIONS?