Diabetes Prediction

Dr. Siadat – Reza Barahmand

Kharazmi University
School of Business
MBA – MIS

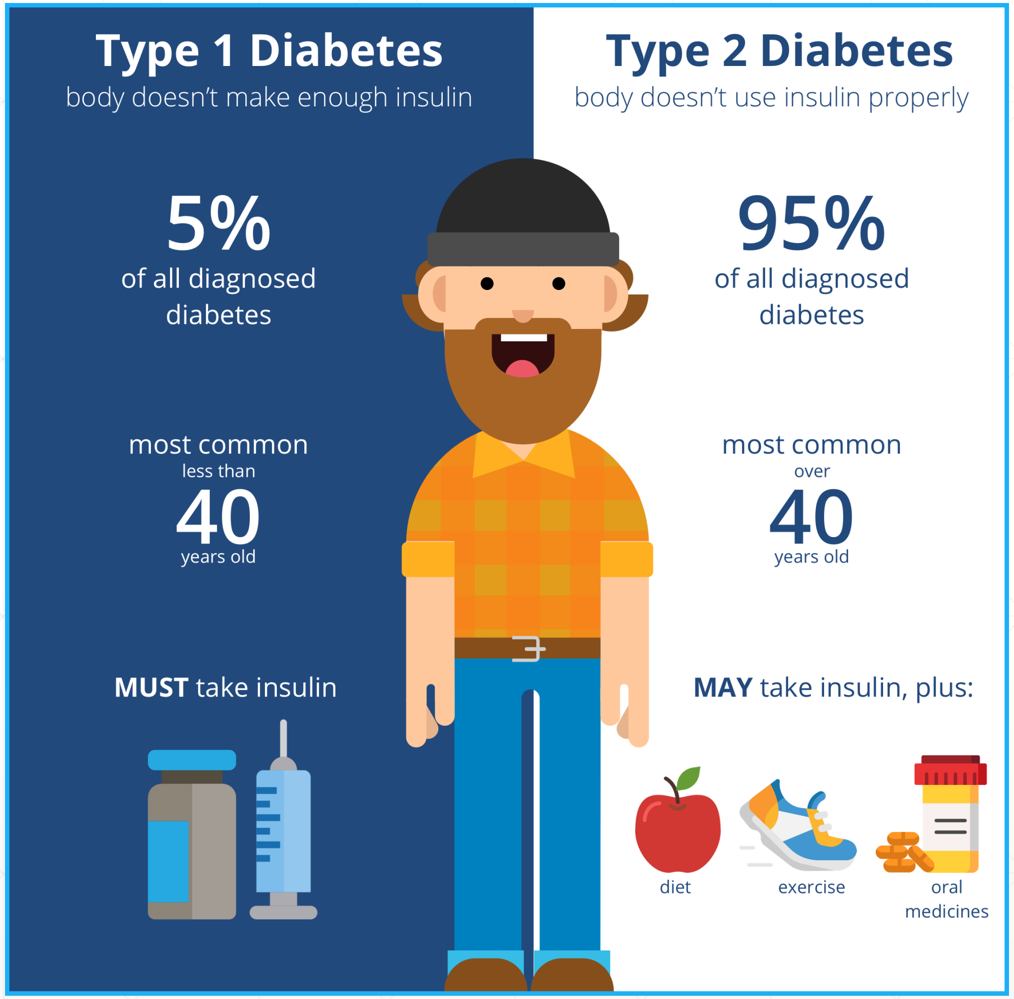
Contents

Introduction

- **Internet of Medical Things**
  - Application of the Internet of Things (IoT) in the medical field
  - Network technologies and it's connection with medical equipment
  - Healthcare IT systems
  - Remote (lacking medical experts)
  - Constant data computation
  - Benefit of using patient records
  - Lower the cost of medical services
  - Delivering feedback to medical staff
  - ML techniques used because of the large amount of data
  - Combination of AI and IoMT is a game changer in this field

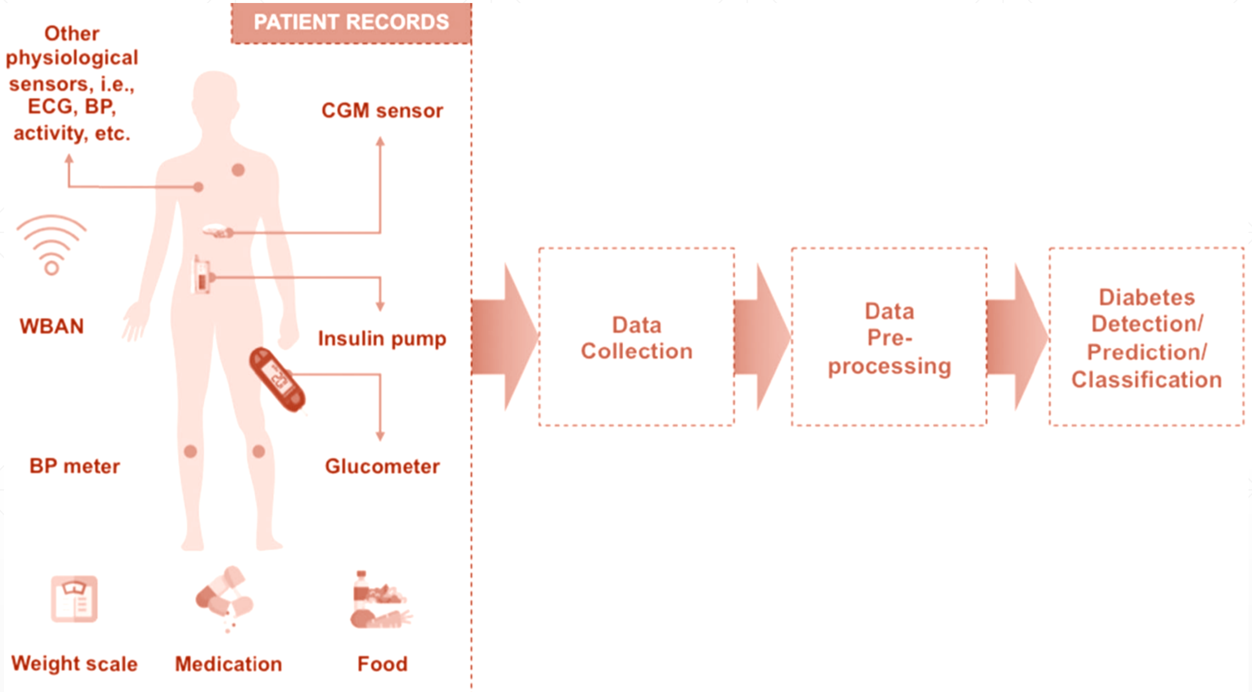
Introduction

- **Diabetes**
  - Chronic illness
  - Develops in 2 situations:
    - Pancreas are not able to generate sufficient insulin
    - Body does not utilize the insulin produced effectively
  - Why people get it?
    - Genetic factors
    - Environmental factors
  - Type 1:
    - Need to inject insulin every day
    - Has no cure
  - Type 2 (our main focus):
    - Blood sugar need to be testes constantly
    - Can be prevented in early stages with healthy diet



Introduction

- **Objectives**
  - Early detection of diabetes
  - Using patient records to accelerate the diagnostic procedure.
  - Using ML and DL to achieve maximum accuracy in prediction.
  - Remote prediction (lacking medical experts)
  - Provide doctors; preliminary diagnosis
  - Feedback doctors about patient records
    - Diet
    - Exercise
    - Blood glucose testing

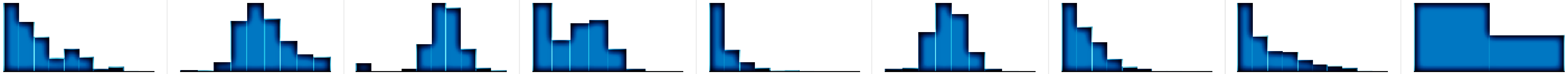


Introduction

- **Pima Indians Diabetes Dataset**
  - National Institute of Diabetes and Digestive and Kidney Diseases
  - All patients here are females at least 21 years old of Pima Indian heritage (768 records)
  - Predictor variables (8 features):
    - Number of pregnancies
    - Glucose
    - Blood pressure
    - Skin Thickness
    - …
  - Target Value:
    - 1 = Has diabetes
    - 0 = Does not have diabetes
  - Data problems
    - Some values inserted as zero that is no possible
    - Data suffers from outliers in some fields
    - Data lacks standardization
    - Imbalanced target

Introduction

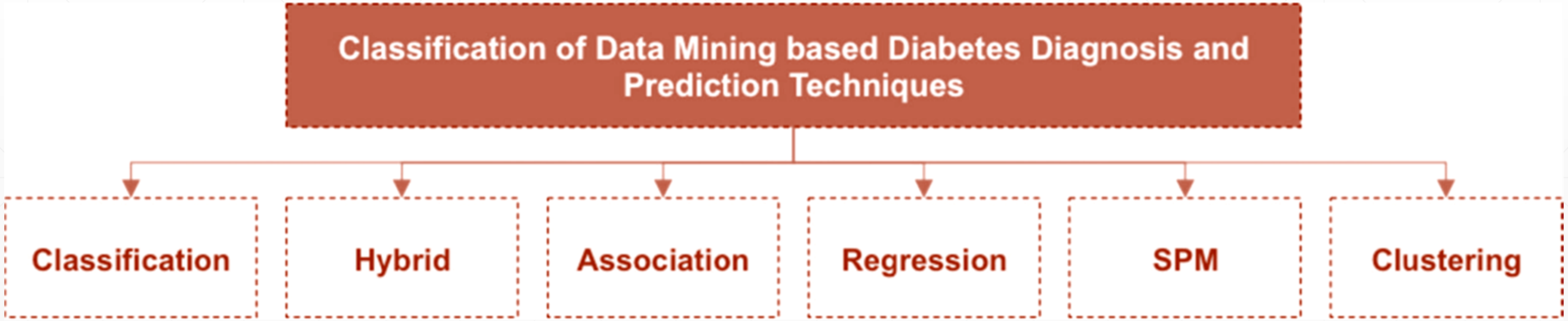
- **Pima Indians Diabetes Dataset**



# Pregnancies	# Glucose	# BloodPressure	# SkinThickness	# Insulin	# BMI	# DiabetesPedigree...	# Age	# Outcome
Number of times pregnant	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Diastolic blood pressure (mm Hg)	Triceps skin fold thickness (mm)	2-Hour serum insulin (mu U/ml)	Body mass index (weight in kg/(height in m)^2)	Diabetes pedigree function	Age (years)	Class variable (0 or 1) 268 of 768 are 1, the others are 0
0 17	0 199	0 122	0 99	0 846	0 67.1	0.08 2.42	21 81	0 1
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0

Methodology

- **Data Mining Based Prediction Techniques**
  - Classification-based:
    - Supervised
    - Data preparation is a plus
  - Regression-based:
    - Statistical
    - Based on relationship between 2 feature
  - Association-based:
    - Extracting frequent pattern and correlations
  - Clustering-based:
    - Unsupervised
    - Base on similarity
  - SPM (Sequential Pattern Mining)
    - Finding patterns, happened orderly.
  - Hybrid
    - Combination of different models
    - Most robust one



Methodology

- **Over All Pipeline (Mine)**

Data Preparation

Data Cleaning:
- Data duplication
- Noisy data
- Outliers
- Missing data

Feature Engineering:
- Standardization
- Feature encoding
- Feature selection
- Feature extraction
- Imbalanced data

Modeling

Data Sampling:
- Stratification
- Cross validation

Modeling:
- Best model selection
- Hyper parameter tuning
- Optimization

Evaluation

Results:
- Max prediction score
- Best evaluation metric
- Prevent overfitting
- Implement the model
- Final test
- Visualizations

Methodology

- **Over All Pipeline (Classic ML Model)**

Data Preparation

Data Cleaning:
- Missing data (Imputation)

Feature Engineering:
- Standardization
- Feature selection (Correlation)

Modeling

Data Sampling:
- 80-20% sampling

Modeling:
- Naïve bayes
- J48 decision tree
- Random forest
- Hyper parameter tuning

Evaluation

Results:
- Max prediction score
- Best evaluation metric
- Implement the model
- Visualizations

Methodology

- **Over All Pipeline (Neural Network)**

Data Preparation

Data Cleaning:

Feature Engineering:
- Neural network
- NN Auto scaling

Modeling

Data Sampling:
- Neural Network

Modeling:
- Quasi-Newton Model
- NN optimization
  - Performance function
  - Learning Rate
  - Numbers of epoch

Evaluation

Results:
- Min error scores
- Implement the model
- Visualizations

Results

- **Confusion Matrix**

		PREDICTED	
		Positive	Negative
ACTUAL	Positive	TRUE POSITIVE	FALSE NEGATIVE
	Negative	FALSE POSITIVE	TRUE NEGATIVE

Results

- **Evaluation Metrics**

$$\textit{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\textit{Specificity} = \frac{TN}{TN + FP}$$

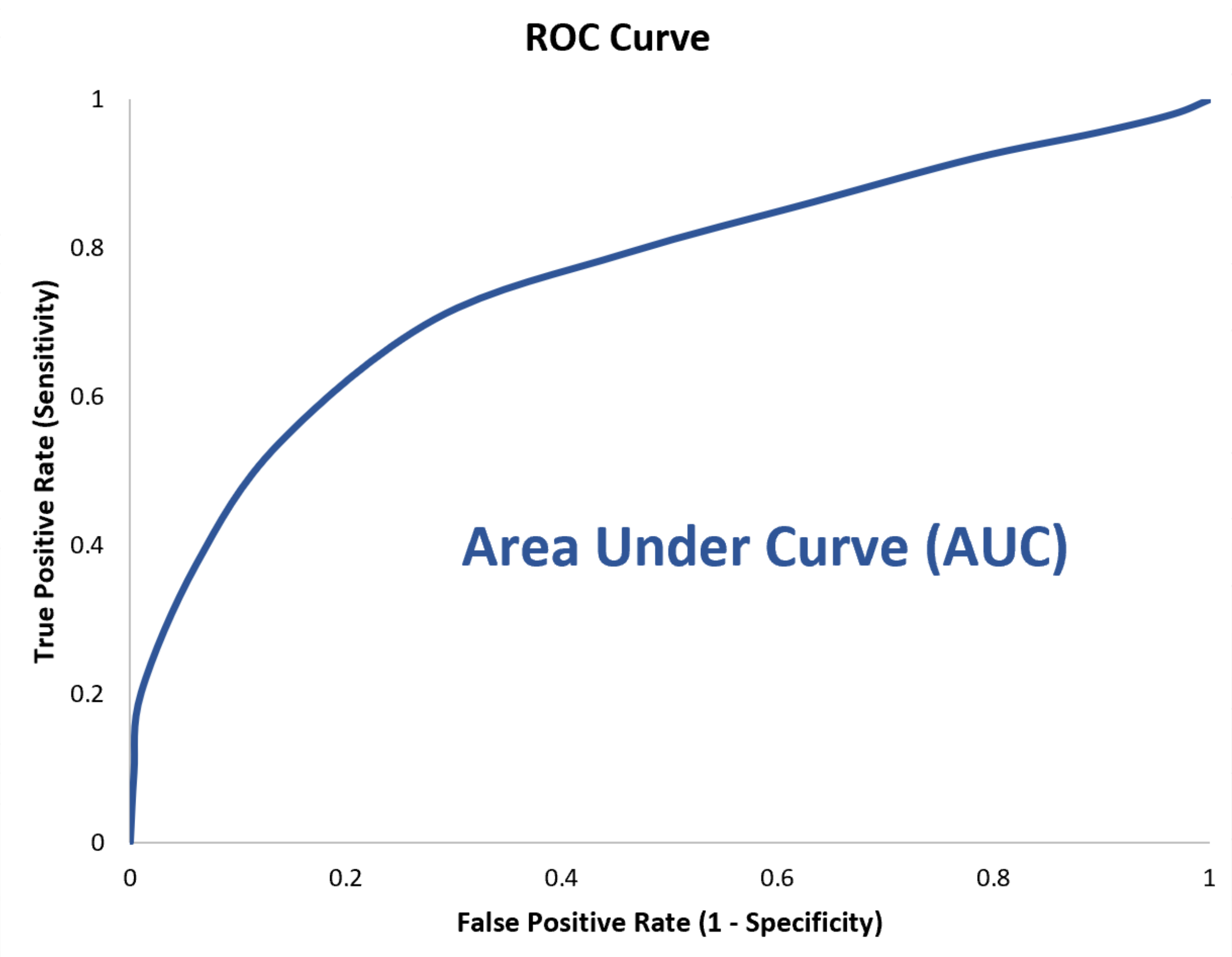
$$\textit{Precision} = \frac{TP}{TP + FP}$$

$$\textit{F - score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

$$\textit{Recall} = \frac{TP}{TP + FN}$$

Results

- **Evaluation Metrics (ROC-AUC)**



Results

- **Error Scores**

Sum Over All The Data Points

Square The Result

$$SS_{Total} = \sum (y_i - \bar{y})^2$$

Sum Squared Total Error

Each Data Point

Mean Value

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

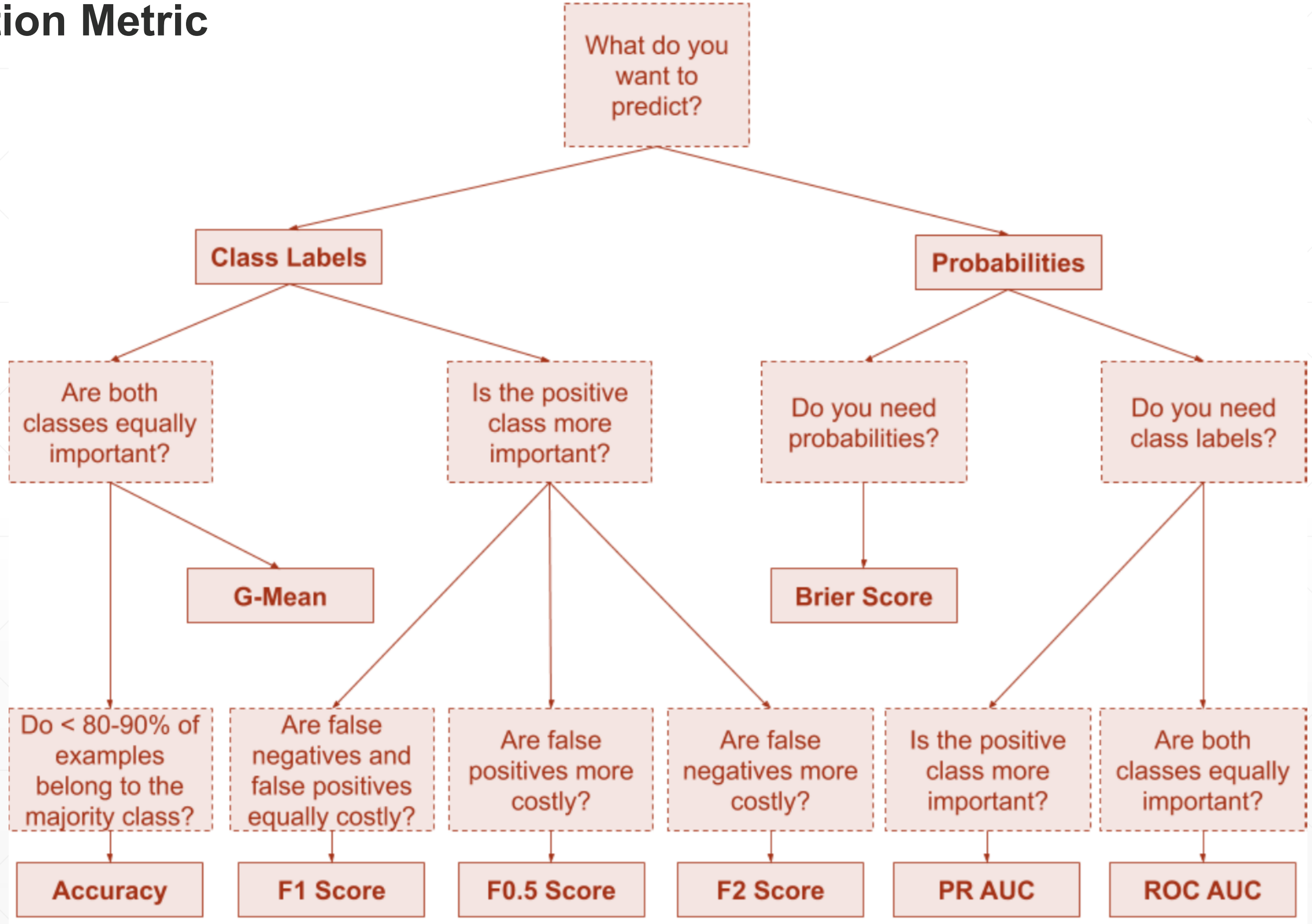
$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

$$\text{WMSE} = \frac{1}{n} \frac{\sum_{i=1}^{1} \text{weights}_i \left(\widehat{\text{predicted}}_i - \text{actual}_i\right)^2}{\sum_{1=1}^{n} \text{weights}_i}$$

Suitable for imbalanced classes

Results

- **Best Evaluation Metric**



Results

- **Classic Models Results (J48 decision tree)**

Table 4 J48 decision tree confusion matrix

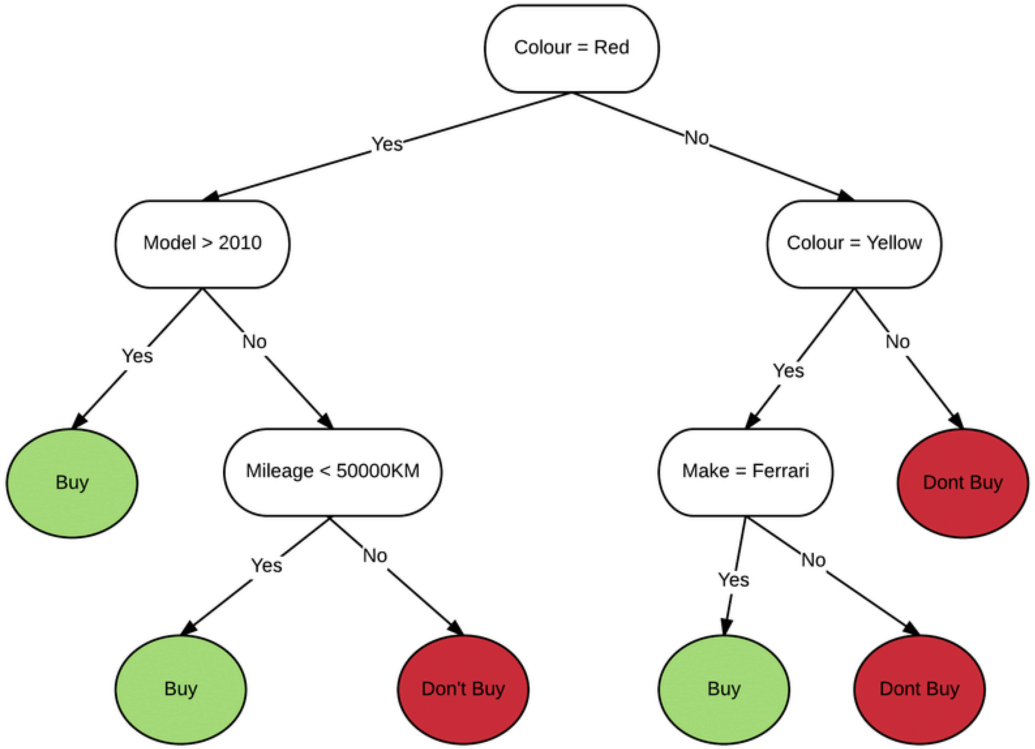
	Actual positive	Actual negative
Predicted positive	107	44
Predicted negative	14	65

Table 5 J48 decision tree confusion matrix with feature selection (3-factor)

	Actual positive	Actual negative
Predicted positive	106	45
Predicted negative	12	67

Table 6 J48 decision tree confusion matrix with feature selection (5-factor)

	Actual positive	Actual negative
Predicted positive	107	44
Predicted negative	12	67



Results

- **Classic Models Results (Random Forest)**

Table 7 Random forest confusion matrix

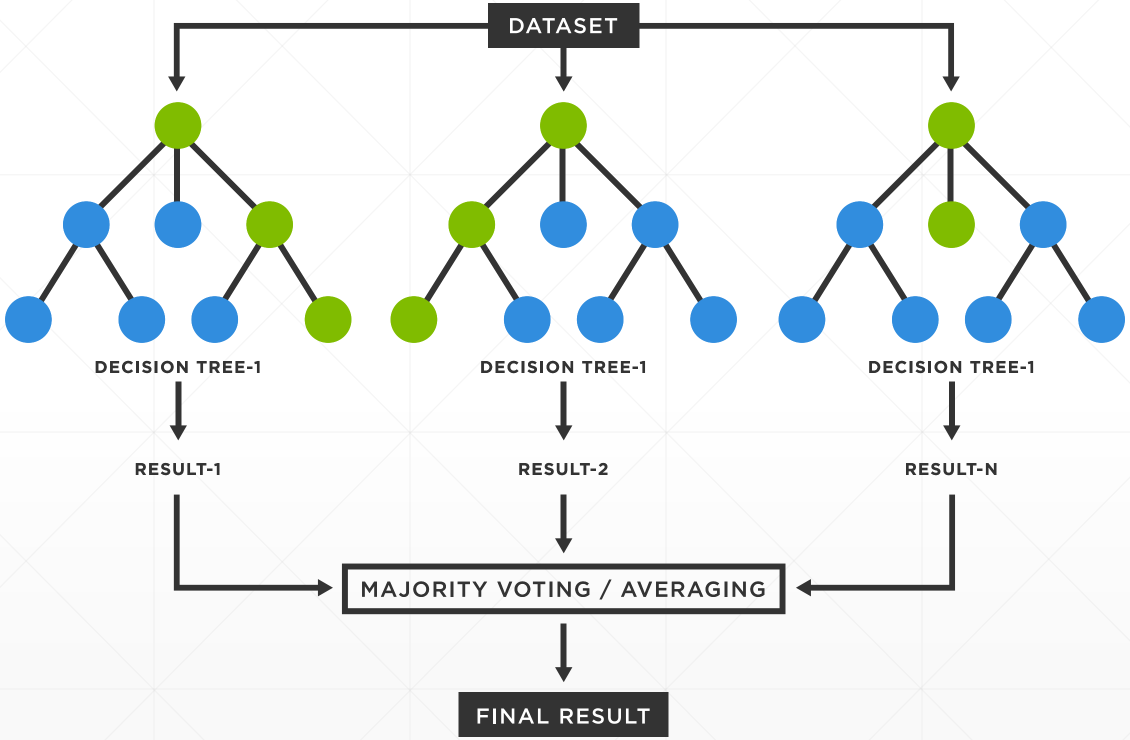
	Actual positive	Actual negative
Predicted positive	136	15
Predicted negative	28	51

Table 8 Random forest confusion matrix with feature selection (3-factor)

	Actual positive	Actual negative
Predicted positive	123	28
Predicted negative	31	48

Table 9 Random forest confusion matrix with feature selection (5-factor)

	Actual positive	Actual negative
Predicted positive	121	30
Predicted negative	30	49



Results

- **Classic Models Results (Naïve Bayes)**

Table 10 Naive Bayes confusion matrix

	Actual positive	Actual negative
Predicted positive	131	29
Predicted negative	20	50

Table 11 Naive Bayes confusion matrix with feature selection (3-factor)

	Actual positive	Actual negative
Predicted positive	133	30
Predicted negative	18	49

Table 12 Naive Bayes confusion matrix with feature selection (5-factor)

	Actual positive	Actual negative
Predicted positive	130	30
Predicted negative	21	49

Likelihood of the Evidence given that the Hypothesis is True

Prior Probability of the Hypothesis

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

Posterior Probability of the Hypothesis given that the Evidence is True

Prior Probability that the evidence is True

Results

- **Classic Models Results (All Models)**

Table 13 Results of all models using the only imputation

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	*F*-score (%)	AUC (%)
J48 decision tree	74.78	70.86	**88.43**	59.63	78.68	78.55
Random forest	**79.57**	**89.40**	81.33	**75.00**	**85.17**	**86.24**
Naïve Bayes	78.67	81.88	86.75	63.29	84.24	84.63

Table 14 Results of all models using feature selection (3-factor)

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	*F*-score (%)	AUC (%)
J48 decision tree	75.22	70.20	**89.83**	59.82	78.81	81.28
Random forest	75.22	**82.12**	80.52	**64.47**	81.31	82.27
Naïve Bayes	**79.13**	81.60	88.08	62.03	**84.71**	**86.15**

Table 15 Results of all models using feature selection (5-factor)

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	*F*-score (%)	AUC (%)
J48 decision tree	75.65	70.86	**89.92**	60.36	79.26	80.84
Random forest	73.91	80.79	79.74	**62.34**	80.26	81.77
Naïve Bayes	**77.83**	**81.25**	86.09	62.03	**83.60**	**84.10**

Results

- **Neural Network Results (All Models)**

Feature	Values
Pregnancies	3.84505
Glucose	120.895
Blood Pressure	69.1055
Skin Thickness	20.5365
Insulin	79.7995
BMI	31.9926
DPF	0.471876
Age	50.2409
Outcome	0.494677295

Example output of NN

Error type	Training	Selection	Testing
Sum squared error	51.7483	38.3264	33.4827
Mean squared	0.112009	0.250499	0.218841
Root mean squared	0.334678	0.500499	0.467805
Normalized squared	0.494779	1.08793	0.966577
Cross entropy error	0.666707	1.75652	1.47763
Minkowski error	64.3526	43.4166	38.1428
Weighted squared	0.434355	1.03511	0.832647

Error table

Results

- **Overall Best Results (Best Model Selected)**

Model	Score
SVM (Mine)	ROC-AUC = 87.6715 %
Random Forest	ROC-AUC = 86.24 %
J48 Decision Tree	ROC-AUC = 81.28 %
Naïve Bayes	ROC-AUC = 86.15 %
Neural Networks	WMSE = 0.434355 (train) – 0.832647 (test)
Maximum Ever	ROC-AUC = 90.12 %

Model	Imbalance	Balanced (SMOTE)
SGD-EN (Stocastic Gradient Decent)	84	85.0041
Logistic Regression	85.9	84.9778
Random Forest	85.7	85.9652
SVM	85.73	87.6715
KNN	86.63	86.9467
Naïve Bayes	0.8467	-
XGBOOST	82.1	0.839518

Future Works

- **Challenges and Recommendations**
  - Data:
    - Availability of relevant accurate and quality data
    - Data collection and sharing
    - Data privacy & security
    - Data integration
    - Data access and storage
  - Data preparation:
    - Appropriate data selection
    - Data cleaning
    - Feature selection and extraction
    - Dimensionality reduction
    - Feature engineering
  - Diagnosis and Prediction Techniques:
    - Generic and universal technique
    - Clinical and public usability
    - Evaluation of existing techniques over new datasets
    - Robust software tools
    - Development a Realtime prediction system
    - Appropriate model selection
    - Integration of models from different domains
    - Higher efficiency and accuracy

Resources

- Victor Chang et al. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. https://link.springer.com/article/10.1007/s00521-022-07049-z. 2022.
- Farrukh Aslam Khan et al. Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review. https://doi.org/10.1109/ACCESS.2021.3059343. 2021.
- Kamlesh Lakhwani et al. Prediction of the Onset of Diabetes Using Artificial Neural Network and Pima Indians Diabetes Dataset. https://doi.org/10.1109/ICRAIE51050.2020.9358308. 2020.
- Pima Indian Diabetes Dataset. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.
- Jason Brownlee. Imbalanced Classification with Python. https://machinelearningmastery.com/imbalanced-classification-with-python\. 2021.
- Jason Brownlee. Data Preparation for Machine Learning. https://machinelearningmastery.com/data-preparation-for-machine-learning/. 2020.
- Reza Barahmand. Jupyter Notebook and Results. https://github.com/rbarahmand/2-years-data-scince-journey/tree/master/docs/myCodes/pima-Indians. 2022.
- IBM CRISP Methodology. https://inseaddataanalytics.github.io/INSEADAnalytics/CRISP_DM.pdf. 2011.

Contents

THANKS !

thanks!

Any questions?

@rbarahmand