THESIS PROJECT

DIABETES PREDICTION USING MULTILAYER PERCEPTRON

Yussup Tumgoyev
ID: 201490584

OUTLINE

1. **Introduction**
    a. Background information
    b. Motivation
    c. Problem statement
    d. Objectives
2. **Literature Review**
3. **Materials and Methods**
4. **Experiments and Results (+ discussion)**
5. **Key contribution, limitations and future work**
6. **Conclusion**

01 BACKGROUND INFORMATION

General introductory information about diabetes

GENERAL INFORMATION

WHAT IS DIABETES?

Diabetes mellitus is a **violation of carbohydrate metabolism** based on:

- **Relative insulin deficiency** - inability of increased concentrations of the hormone to normalize blood sugar levels
- **Insulin resistance** - decrease in the susceptibility of cells to the action of insulin

327,000,000,000

$1 out of every $4 in US health care costs is spent on caring for people with diabetes (cdc.gov)

from 108 millions to

422,000,000

According to researchers' forecasts, this statistics will increase to **700 million** in **20 years.**

1,500,000

WHO ranks diabetes to be **the top nine causes of death** among all the diseases

iDala
virtual healthcare

Seitzhan Sypabekov

CEO, iDala Learn more

Yussup Tumgoyev

CTO, iDala Learn more

Amanzhol Shungeyev

CIO, iDala Learn more

www.idala.asia

PROBLEM & RQ

PROBLEM

The dire need for the **tool** that can help with **early diabetes detection** to prevent serious **medical** and **financial complications** of this disease

RESEARCH QUESTION

Does **simple neural network** models like MLP has a **potential to predict** the risk of the diabetes mellitus type 2?

02

LITERATURE REVIEW

Study of the previously made researchers in this field

LITERATURE REVIEW

	Algorithms	Results
Swapna G. et al.	CNN, LSTM and **Support Vector Machines** (2018)	**95.7% accuracy**
Ashok k. et al.	SVM, **Logistic Regression**, ANN, DT, and KNN (2018)	**78% accuracy**
Nesreen L. et al.	**Artificial Neural Network** (2018)	**87% accuracy**
Deepti S. et al.	**Naive Bayes**, SVM, and DT (2018)	**76% accuracy**
Ram D. et al.	**Logistic Regression** (2021)	**78% accuracy**
Sidong W. et al	**Deep Neural Network**, SVM, DT, and NB (2018)	**78% accuracy**
Aiswarya I.	**Naive Bayes**(2015)	**80% accuracy**
Quan Z. et al.	**Decision Tree** (2015)	**77% accuracy**
Vijayan V. et al.	**Gradient Boosting**, LR, and NB (2019)	**80.7% accuracy**

LITERATURE REVIEW

	Algorithms	Results
Deepti S. et al.	DT, SVM, and **Naive Bayes** (2018)	**0.82 AUC**
Kamrul H. et al.	**Multilayer Perceptron** (2020)	**0.90 AUC**

The MLP model proposed by **Kamrul H.** were taken as a **baseline**

MULTILAYER PERCEPTRON WAS DECIDED TO BE RESEARCHED WITH AREA UNDER THE ROC CURVE AS A PERFORMANCE METRIC

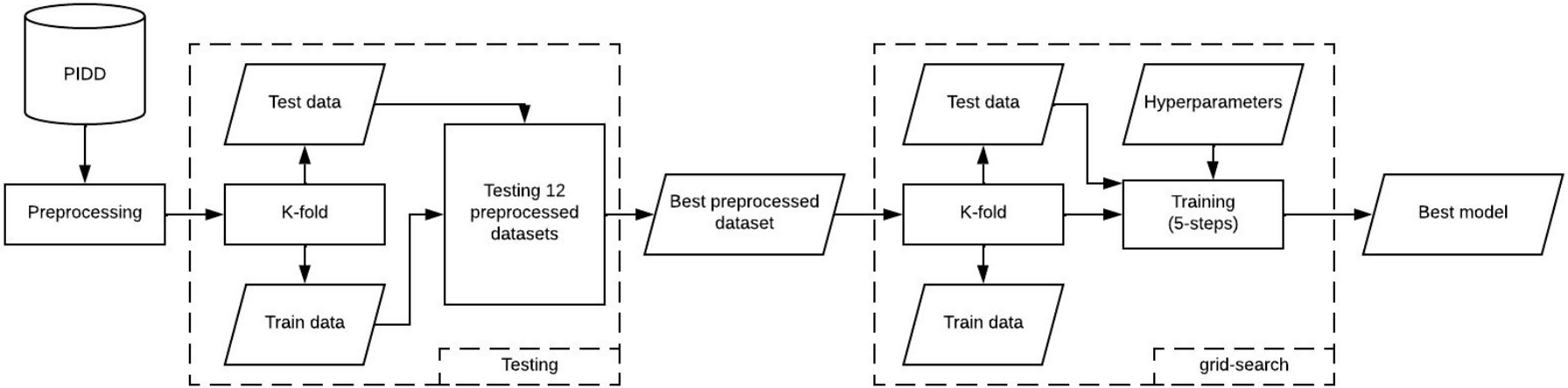
Proposed idea

03

MATERIALS AND METHODS

Materials and methods used in this study

PROPOSED FRAMEWORK



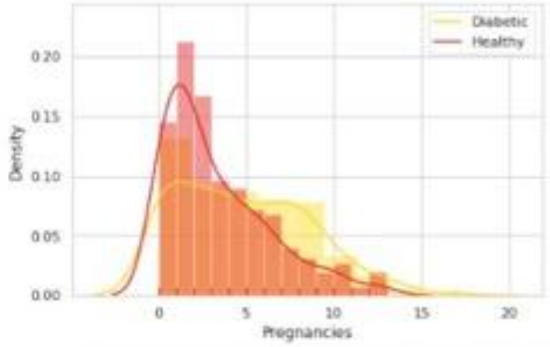
PID DESCRIPTION

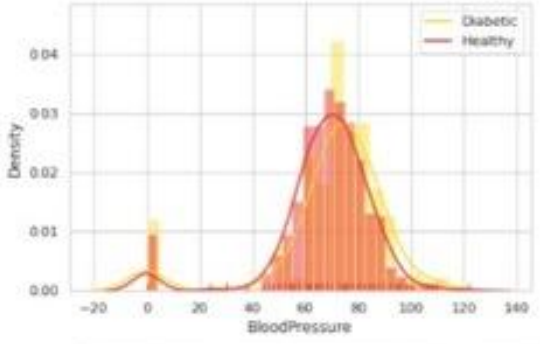
#	Atribute	Mean ± stdev	Range
1	Pregnancies	3.85 ± 3.37	0 - 17
2	Glucose	120.90 ± 31.97	0 - 199
3	BloodPressure	69.11 ± 19.36	0 - 122
4	SkinThickness	20.54 ± 15.95	0 - 99
5	Insulin	79.81 ± 115.24	0 - 846
6	BMI	32.00 ± 7.88	0 - 67.1
7	DiabetesPedigreeFunction	0.47 ± 0.33	0.078 - 2.42
8	Age	33.24 ± 11.76	21 - 81

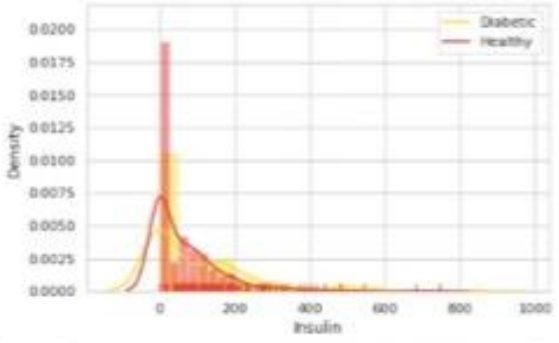
DATASET LIMITATION

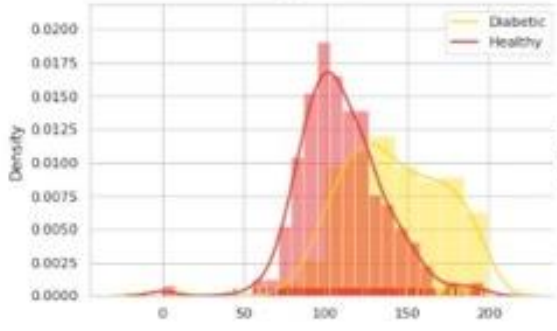
PIMA INDIAN DATASET

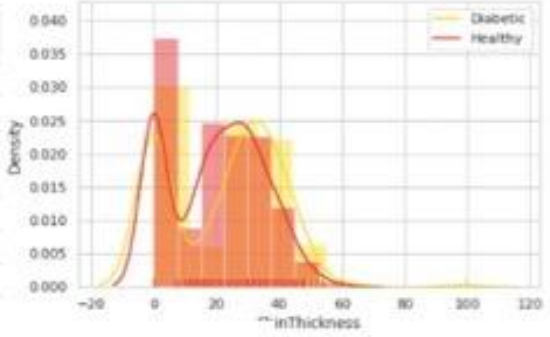
- **Small dataset** of 768 instances
- **Missing values**: 51% of the data
- **Outliers**: 126/768 instances
- **Imbalanced class labels**: 268(-)/500(+)
- **One gender**

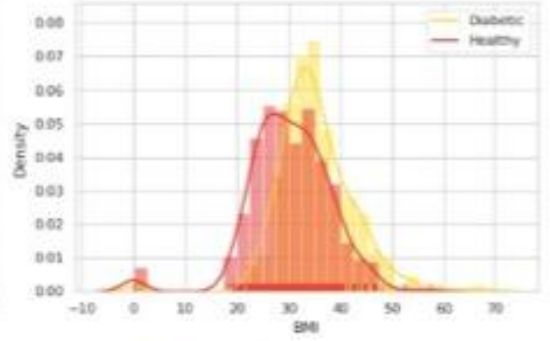

Diabetic
Healthy
Density
Pregnancies

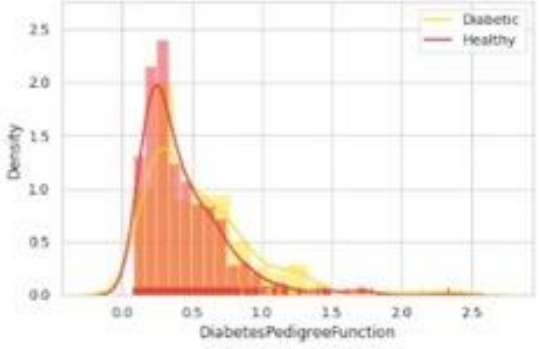

Diabetic
Healthy
Density
BloodPressure

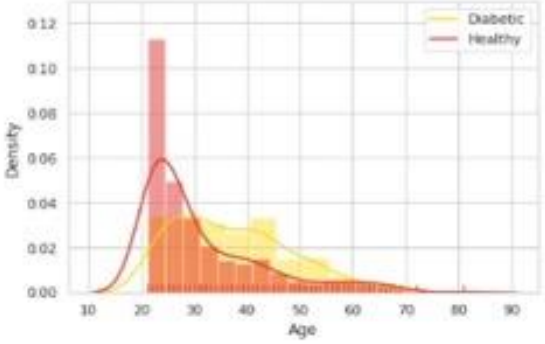

Diabetic
Healthy
Density
Insulin


Diabetic
Healthy
Density


Diabetic
Healthy
Density
inThickness


Diabetic
Healthy
Density
BMI


Diabetic
Healthy
Density
DiabetesPedigreeFunction


Diabetic
Healthy
Density
Age

DATA PREPROCESSING

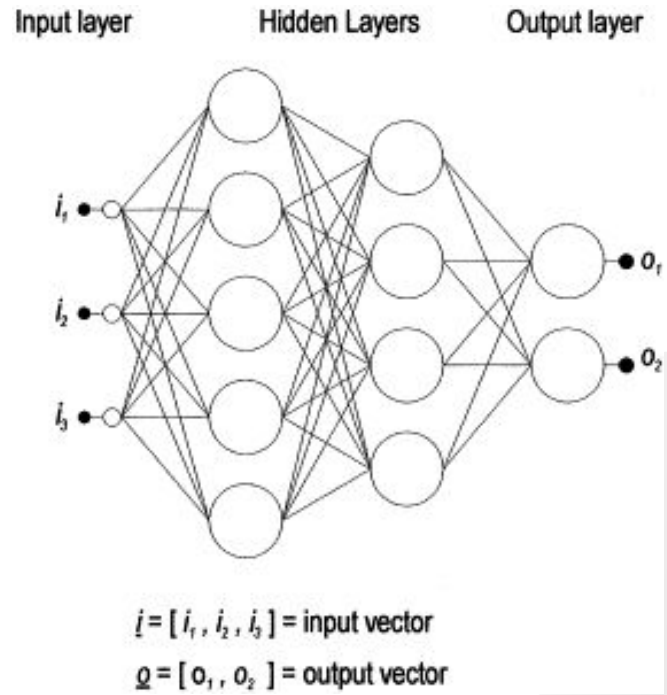
1. Outlier rejection
2. Missing values imputation
3. Normalization
4. Feature selection (PCA / ICA)

Outlier rejection (O)				(O) + Missing values imputation (M)				(O)+(M)+Z standardization (Z)			
PCA		ICA		PCA		ICA		PCA		ICA	
4	6	4	6	4	6	4	6	4	6	4	6
dataset 1	dataset 2	dataset 3	dataset 4	dataset 5	dataset 6	dataset 7	dataset 8	dataset 9	dataset 10	dataset 11	dataset 12

K-FOLD CROSS-VALIDATION

MULTILAYER PERCEPTRON

- 3 types of layers
- **Interconnected** nodes
- **Activation function** on each neuron
- **Back propagation** algorithm

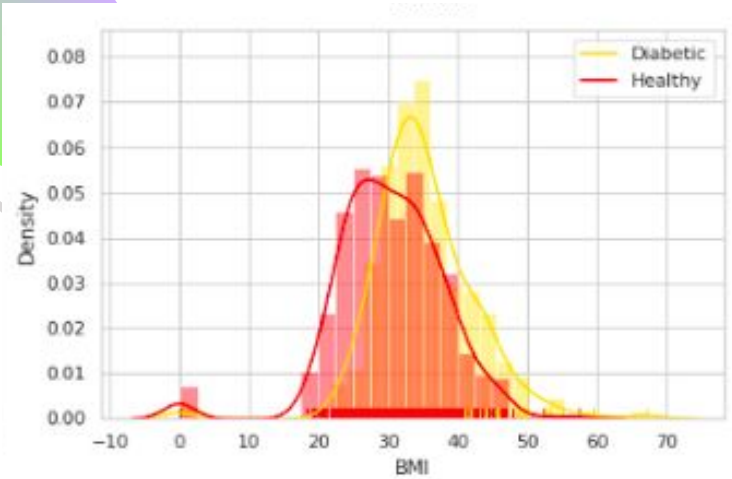


04

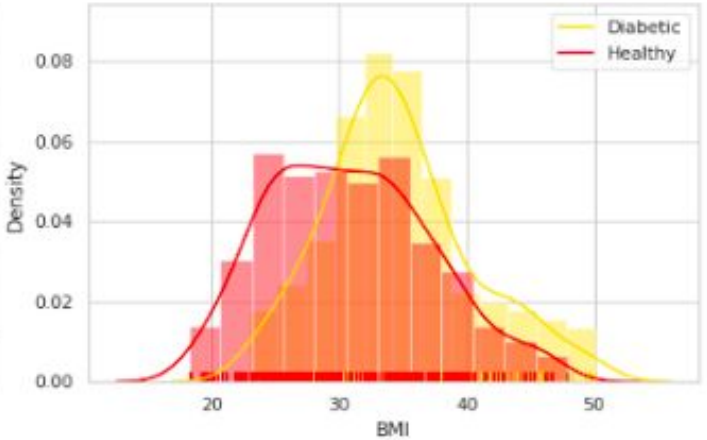
EXPERIMENTS AND RESULTS

Experiments performed to find the best MLP model and results

DATA PREPROCESSING



BEFORE



AFTER

DATA PREPROCESSING

Feature name	# outliers
Pregnancies	4
Glucose	5
BloodPresure	45
SkinThickness	1
Insulin	34
BMI	19
PedigreeFunctio	29
Age	9
Total: 129 rows with outliers (17% of all data)	

Figure 5-3: The number of outliers in each row

Feature name	# missing values
Pregnancies	0
Glucose	0
BloodPresure	0
SkinThickness	179
Insulin	307
BMI	0
PedigreeFunctio	0
Age	0
Total	486 items (9.51% of all data)

Table 5.1: Missing values after outlier rejection

SEVEN-STEP METHOD

1. The best dataset
2. The best architecture
3. Alternative PCA feature selection
4. Hyperparameters tuning
5. Alternative data preprocessing
6. More experiments
7. Model validation

BEST DATASET

Results	Outlier rejection (O)				(O) + Missing values imputation (M)				(O)+(M)+Z standardization (Z)			
	PCA		ICA		PCA		ICA		PCA		ICA	
	4	6	4	6	4	6	4	6	4	6	4	6
	dataset 1	dataset 2	dataset 3	dataset 4	dataset 5	dataset 6	dataset 7	dataset 8	dataset 9	dataset 10	dataset 11	dataset 12
This literature	0.800	0.818	0.797	0.793	0.802	0.821	0.796	0.801	0.832	0.889	0.801	0.802
Baseline model	0.738	0.770	0.787	0.829	0.846	0.874	0.901	0.887	0.890	0.881	0.889	0.885
Difference	0.062	0.048	0.01	-0.036	-0.044	-0.053	-0.105	-0.086	-0.058	0.008	-0.088	-0.083

- **Architecture is the same** as baseline
- **12 datasets** were the result of combination of the preprocessing techniques
- Replicated MLP worse by **0.04 AUC** than baseline (PCA **0.006 AUC** worse, ICA **0.065 AUC**)
- **Controversial results**

BEST ARCHITECTURE

# of hidden layers	Neurons on each layer	Performance (AUC)	Loss curve behavior
2	8-4	0.889	good fit
3	4-32-16	0.887	good fit
2	16-4	0.887	good fit
2	4-8	0.886	good fit
2	8-8	0.886	good fit
3	8-16-128	0.886	good fit
2	8-64-64	0.886	good fit
2	8-32	0.885	good fit
3	4-128	0.885	good fit
3	16-64-16	0.885	good fit
3	4-32-32	0.884	good fit
3	16-16-64	0.884	good fit
2	16-32	0.884	good fit
1	4	0.883	good fit
1	8	0.883	good fit

Table 5.3: Step 2: The performance of top 15 architectures with different combinations of hidden layers and neurons on each layer.

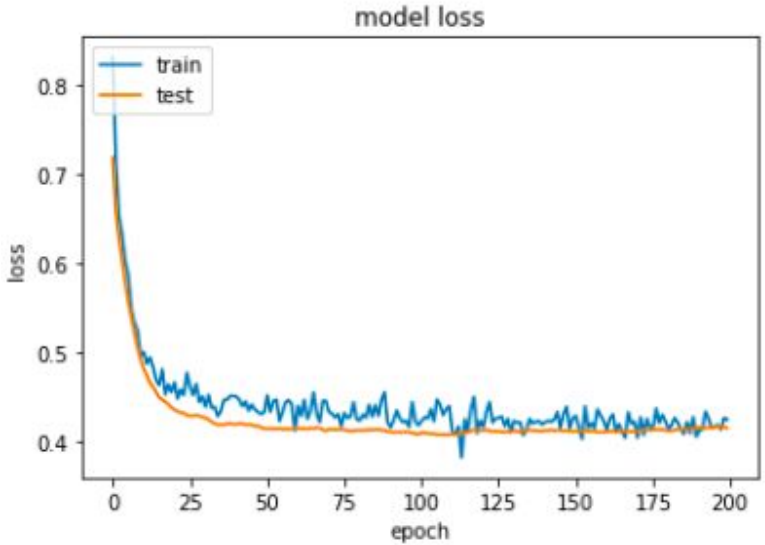


Figure 5-4: The good fitted loss learning curve

BEST ARCHITECTURE

# of hidden layers	Neurons on each layer	Performance (AUC)	Loss curve behavior
2	8-4	0.889	good fit
3	4-32-16	0.887	good fit
2	16-4	0.887	good fit
2	4-8	0.886	good fit
2	8-8	0.886	good fit
3	8-16-128	0.886	good fit
2	8-64-64	0.886	good fit
2	8-32	0.885	good fit
3	4-128	0.885	good fit
3	16-64-16	0.885	good fit
3	4-32-32	0.884	good fit
3	16-16-64	0.884	good fit
2	16-32	0.884	good fit
1	4	0.883	good fit
1	8	0.883	good fit

Table 5.3: Step 2: The performance of top 15 architectures with different combinations of hidden layers and neurons on each layer.

- **1 to 3 hidden layers (hl)**
- **4, 8, 16, 32, 64, or 128 neurons** on each layer
- Best performance of **0.889 (± 0.02) AUC**
- **6 architectures of 3 hl** and **8 have less than 3 hl (?)**

ALTERNATIVE PCA

# of layers	Neurons on each layer	Performance (AUC) PCA 3	Performance (AUC) PCA 5	Performance (AUC) PCA 6	Performance (AUC) PCA 7	Loss curve behavior
2	8-4	0.828	0.848	0.889	0.883	good fit
3	4-32-16	0.828	0.829	0.887	0.873	good fit
2	16-4	0.832	0.841	0.887	0.888	good fit
2	4-8	0.831	0.817	0.886	0.875	good fit
2	8-8	0.837	0.839	0.886	0.882	good fit
3	8-16-128	0.831	0.838	0.886	0.879	good fit
3	8-64-64	0.821	0.828	0.886	0.877	good fit
2	8-32	0.834	0.839	0.885	0.886	good fit
2	4-128	0.834	0.842	0.885	0.887	good fit
3	16-64-16	0.838	0.836	0.885	0.876	good fit
3	4-32-32	0.827	0.820	0.884	0.874	good fit
3	16-16-64	0.837	0.831	0.884	0.887	good fit
2	16-32	0.841	0.843	0.884	0.891	good fit
1	4	0.836	0.828	0.883	0.884	good fit
1	8	0.840	0.840	0.891	0.888	good fit

Table 5.4: Step 3: The performance of top 15 architectures with different number of hidden layers and primcipal components of PCA.

- **3, 5, 6, and 7 PC**
- **PCA 6** is still the best feature selection
- **8 architectures selected** with 0.886 AUC or above

HYPERPARAMS TUNING

# of layers	Neurons on each layer	Performance (AUC) PCA 3	Batch Size	Learning Rate	Dropout	Loss curve behavior
1	8	0.900	16	0.001	60%	good fit
2	16-32	0.892	16	0.001	60%	good fit
3	16-16-64	0.838	8	0.001	60%	good fit
2	4-128	0.884	4	0.01	60%	good fit
2	4-8	0.887	16	0.001	70%	good fit
2	16-4	0.887	8	0.0001	50%	good fit
3	4-32-16	0.888	16	0.01	70%	good fit
2	8-4	0.886	32	0.001	50%	good fit

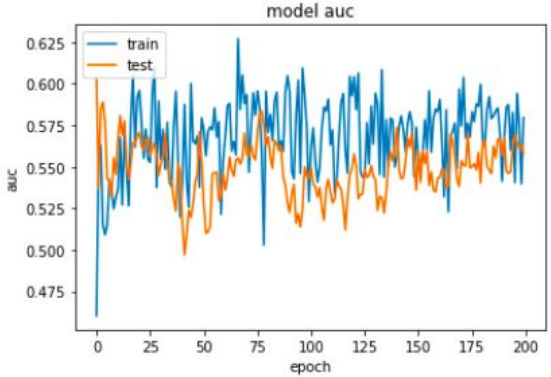
Table 5.5: Best hyperparameters for the 8 architectures selected

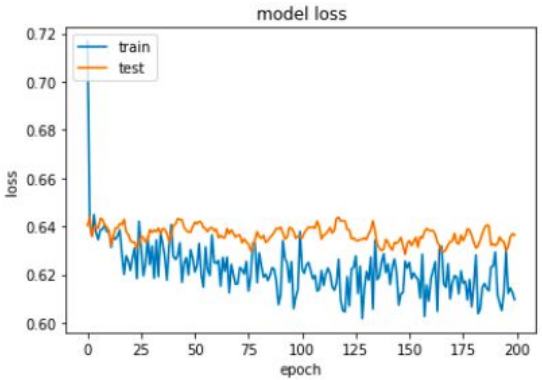
- Batch size [4, 8, ,16, 32, 64, 128]
- Learning rate [0.0001, 0,001, 0.01, 0.1]
- Dropout neurons [0.5, 0.6, 0.7]
- 576 architectures tested
- 0.90 AUC performance

ALTERNATIVE PREPROCESSING

Architecture	Feature selection technique	# of features selected	AUC	Loss curve behavior
1 hidden layer with 8 neurons	information gain	4	0.789	fits
	information gain	5	0.824	fits
	fisher's score	4	0.640	unrepresentative
	fisher's score	5	0.712	unrepresentative
	chi-squared	4	0.643	unrepresentative
	chi-squared	5	0.545	unrepresentative

Table 5.6: The results for best architecture with alterantive feature selection techniques





ALTERNATIVE PREPROCESSING

Architecture	Data Scaler	AUC	Loss curve behavior
1 hidden layer with 8 neurons	StandardScaler	0.545	unrepresentative
	MaxAbsScaler	1.000	unrepresentative
	MinMaxScaler	1.000	unrepresentative
	RobustScaler	0.606	unrepresentative
	Normalizer	1.000	unrepresentative

Table 5.7: Step 5: The results for (O + M) PIDD scaling with sklearn.preprocessing module.

- **Poor results** for chi-squared and fisher's score
- **Huge overfitting** for data scalers when applying with PCA

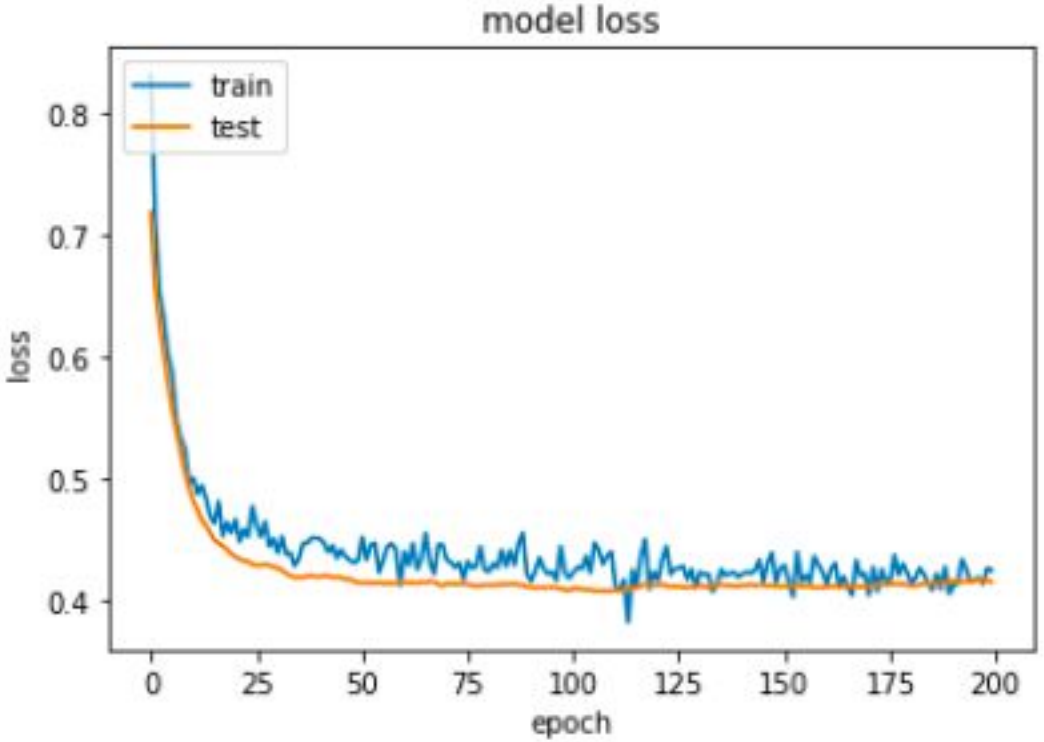
MORE EXPERIMENTS

Architecture	Optimizer	AUC	Loss curve behavior
1 hidden layer with 8 neurons	SGD	0.875	good fit
	RMSprop	0.889	good fit
	Adam	0.897	good fit
	Adadelta	0.561	unrepresentative
	Adagrad	0.624	unrepresentative
	Adamax	0.881	good fit
	Nadam	0.887	good fit
	Ftrl	0.854	unrepresentative

Table 5.8: Step 6: Applying various optimizers with the best MLP model

- **No improvements** with dynamic learning rate
- RMSprop, Adamax, and Nadam optimizers show a **competitive results**
- Best architecture with **50 epochs**

MORE EXPERIMENTS



MODEL VALIDATION

Results	Outlier rejection (O)				(O) + Missing values imputation (M)				(O)+(M)+Z standardization (Z)			
	PCA		ICA		PCA		ICA		PCA		ICA	
	4	6	4	6	4	6	4	6	4	6	4	6
	dataset 1	dataset 2	dataset 3	dataset 4	dataset 5	dataset 6	dataset 7	dataset 8	dataset 9	dataset 10	dataset 11	dataset 12
Reolicated model (R)	0.800	0.818	0.797	0.793	0.802	0.821	0.796	0.801	0.832	0.889	0.801	0.802
Baseline model (Ba)	0.738	0.770	0.787	0.829	0.846	0.874	0.901	0.887	0.890	0.881	0.889	0.885
Best model (Be)	0.805	0.829	0.805	0.833	0.807	0.832	0.804	0.834	0.841	0.901	0.807	0.841
Difference (Be-R)	0.005	0.011	0.008	0.04	0.005	0.011	0.008	0.033	0.009	0.012	0.006	0.039
Difference (Be-Ba)	0.067	0.059	0.018	0.004	-0.039	-0.042	-0.097	-0.053	-0.049	0.020	-0.082	-0.044
Difference (R-Ba)	0.062	0.048	0.01	-0.036	-0.044	-0.053	-0.105	-0.086	-0.058	0.008	-0.088	-0.083

Table 5.9: Step 7: The results for the Best MLP model versus Baseline and Replicated

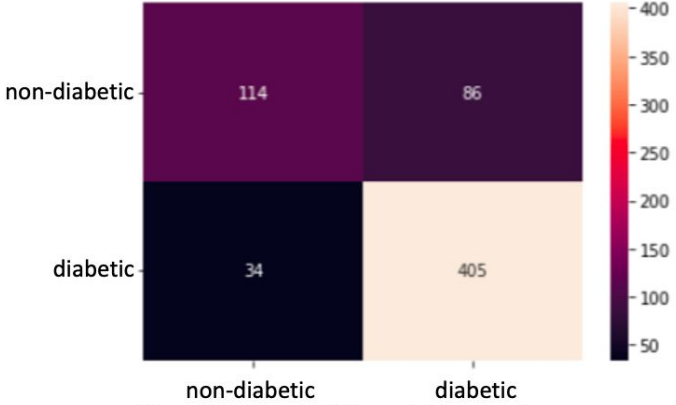


Figure 5-6: Step 7: The confusion matrix

	Baseline model (BA)	Model proposed in this work (BE)
Number of hidden layers	3	1
Total number of neurons	144	8
Training time	174 sec	31 sec
Number of trainamble parameters	5425	65
Memory usage	21.95 kb	0.26 kb
Performance	0.90 AUC	0.90 AUC

Table 5.10: Step 7: The comparison table of the BA versus BE

05

KEY CONTRIBUTION, LIMITATIONS, AND FUTURE WORK

Almost the end of the thesis presentation

KEY CONTRIBUTIONS

- **Broad research on Multilayer Perceptron** of 1-3 hidden layers
- **Lighter MLP model** that is **faster** and memory **efficient**
- **Dataset with MLP performance based on**: num of hidden layers, neurons, activation functions, batch sizes, optimizers, learning rates, dropout layers, feature selection techniques, loss functions

1st HL	2nd HL	3rd HL	AUC	Learning curve	Batch Size	Learning Rate	Dropout	Feature selection	Epochs	Optimizer	Loss
4			88.54	good fit	8	0.001	0.6	PCA_6	200	Adam	Binary Cross-Entropy
8			88.29	good fit	8	0.001	0.6	PCA_6	200	Adam	Binary Cross-Entropy
16			88.11	good fit	8	0.001	0.6	PCA_6	200	Adam	Binary Cross-Entropy
32			87.87	good fit	8	0.001	0.6	PCA_6	200	Adam	Binary Cross-Entropy
64			87.85	good fit	8	0.001	0.6	PCA_6	200	Adam	Binary Cross-Entropy
128			87.58	good fit	8	0.001	0.6	PCA_6	200	Adam	Binary Cross-Entropy
4	4		87.15	good fit	8	0.001	0.6	PCA_6	200	Adam	Binary Cross-Entropy
4	8		88.6	good fit	8	0.001	0.6	PCA_6	200	Adam	Binary Cross-Entropy
4	16		87.19	good fit	8	0.001	0.6	PCA_6	200	Adam	Binary Cross-Entropy

LIMITATIONS & FUTURE WORK

LIMITATIONS:

- The data **normalization and PCA feature selection** applied to the entire dataset
- The work covers the experiments on **1 to 3 hidden layers with 4 to 128 neurons** on each layer
- **Untapped potential** of optimizers and dynamic learning rate
- **Only MLP** was tested
- The data provided by **PIDD seems not enough**
- No assistance from a qualified **medical specialist**

LIMITATIONS & FUTURE WORK

FUTURE WORK:

- Normalization and PCA feature selection to the training dataset
- Deeper MLP model
- Use more powerful hardware
- Other NN models like ResNet
- Use other dataset
- Assistance from a qualified medical specialist

06

CONCLUSION

Summary of the entire study

THE END OF PRESENTATION

THANK YOU!

Do you have any questions?

Feel free to ask

yussup.tumgoyev@nu.edu.kz

IMPORTANT INFORMATION

- Athletes with diabetes should be **screened for complications associated with the disease**.
- Combine **regular aerobic exercise** with **resistance exercise.**
- Insulin-dependent athletes should closely **monitor** their **blood sugar levels before, during**, and **after exercise**.
    - Categorically **not recommended** to **make physical exercise** at **significant hyperglycemia**, since physical activity can **paradoxically aggravate hyperglycemia** and lead to **ketoacidosis.**

RESEARCHES LIMITATIONS

REVIEWED STUDIES

- Use of this low-quality dataset
- Applied solution for the filling of empty values

They fill in the null values with the mean values of all other existing data.

  - Glucose: 5
  - Blood Pressure: 35
  - Skin Thickness: 227
  - Insulin: 374
  - BMI: 11

ALTERNATIVE DATASET

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 167 (2020) 706–716

Procedia
Computer Science

www.elsevier.com/locate/procedia

DD2019

Diabetes Dataset 2019

International Conference on Computational Intelligence and Data Science (ICCIDS 2019)

Prediction of Type 2 Diabetes using Machine Learning Classification Methods

Neha Prerna Tigga[a], Shruti Garg[a,*]

This dataset was collected by Neha Prerna Tigga and Dr. Shruti Garg of the Department of Computer Science and Engineering, Birla Institute of Technology - Mesra

RQ & HYPOTHESIS

RESEARCH QUESTION

Are BMI, regular medicine, blood pressure and stress level the most important parameters for T2DM risk prediction?

HYPOTHESIS

BMI, regular medicine, blood pressure and stress level are the most important parameters for T2DM risk prediction.

THE MAIN GOAL OF THIS STUDY IS TO DEVELOP A MACHINE LEARNING MODEL FOR PREDICTING THE RISK OF TYPE 2 DIABETES WITH AN ACCURACY ABOVE 95%.

Objective

SPECIFIC AIMS

- Approval from the **NU Institutional Ethics Committee**
- Dataset search:
  - Datasets with the diagnosis of **type 2 diabetes**
  - General **initial assessment** of the found dataset
- Dataset preprocessing:
  - **Statistical analysis, distribution checking, finding patterns, assumptions testing**
  - Removing the **outliers**, restoring the **null values** and make **normalization of data frame**
- Final model and it's validation
  - Building the model and identification of the algorithms with **highest accuracy** and it's **feature importance**
  - Validation by **Pima Indian Dataset**

03

EXPERIMENTAL PLAN

Basic steps for completing a study

DD2019 DESCRIPTION

	#	Options
Age	1	Less than 40 / 40-49 / 50-59 / 60 or older
Gender	2	Male / Female
Family history with diabetes	3	Yes / No
Diagnosed with high blood pressure	4	Yes / No
Physical Activity	5	None / Less than 0,5 hr/ More than 0,5 hr/ 1+ hr
BMI	6	Numeric
Smoking	7	Yes / No
Alcohol consumption	8	Yes / No
Sleep duration in hours	9	Numeric
Sound sleep duration in hours	10	Numeric
Regular medicine intake	11	Yes / No
Junk food consumption	12	Occasionally / Often / Very often / Always
Stress level	13	Not at all / Sometimes / Very often / Always
Blood pressure level	14	High / Normal / Low
Number of pregnancies	15	Numeric
Gestational diabetes	16	Yes / No
Urination frequency	17	Not much / Quite often
Diabetic (result)	18	Yes - 266 / No - 685

PID DESCRIPTION

	#	Options
Age (years)	1	Numeric
Pregnancies	2	Numeric
Concentration of plasma glucose (mg/dL)	3	Numeric
Diastolic blood pressure (mm Hg)	4	Numeric
Triceps skin fold thickness (mm)	5	Numeric
2-Hour serum insulin (mu U/ml)	6	Numeric
BMI (kg/m^2)	7	Numeric
A pedigree function for diabetes	8	Numeric
Outcome	9	Yes - 268 / No - 500

DD2019 PREPROCESSING

- Identification of **missing values**
  - BMI: 8
  - Pregnancies: 42
  - Pdiabetes: 1
  - Diabetic: 1

  Result: from 952 to 905 samples
- Conversion of **categorical value** to **nominal** (changing the text data to digital)

  Result: from 17 to 32 parameters

DD2019 PREPROCESSING

- Identification of **outliers**

  Result: From 905 to 872 samples

- **Normalization** of dataset

  Database **normalization** is the **process of structuring a database**, in accordance with a series of so-called normal forms in order to **reduce data redundancy and improve data integrity**.

- The **division of dataset into test and train subsets** with ration **20% and 80%** respectively.

PID PREPROCESSING

- Identification of **zero values**
  - Glucose: 5
  - Blood Pressure: 35
  - Skin Thickness: 227
  - Insulin: 374
  - BMI: 11
- The zero values of **Glucose, Blood Pressure and BMI** were removed.

  Result: From 768 to 724 samples

PID PREPROCESSING

- Identification of **outliers**
  - Pregnancies: 3
  - Glucose: 13
  - Insulin: 28
  - BMI: 6
  - Diabetes Pedigree Function: 27
  - Age: 6

  Result: From 724 to 640 samples
- **Filling the zero values** of **Skin Thickness and Insulin** parameters by **Iterative imputation**
- **Normalization** of dataset **(Canceled)**

DATA PROCESSING

1. **LOGISTIC REGRESSION**

   It is a predictive analysis algorithm and based on the concept of probability

2. **SUPPORT VECTOR MACHINE**

   Classifies data points by an appropriate hyperplane in a multidimensional space

3. **K-NEAREST NEIGHBOR**

   Assumes the similarity between the new and available data and put the new data into the category that is most similar

DATA PROCESSING

4 **DECISION TREE**

Works on the principle of decision making. It can be described in form of tree and provides high accuracy and stability

5 **RANDOM FOREST**

Creates multiple decision tree from randomly selected subset of training dataset, then aggregates the votes from different decision trees to decide the final class of test objects

6 **NAIVE BAYES**

Probabilistic machine learning algorithm based on Bayes theorem described in probability

DATA PROCESSING

7 L2 REGULARIZATION (RIDGE CLASSIFIER)

In supervised machine learning, the ML models get trained training data and there are the possibilities that the model performs accurately on training data but fails to perform well on test data.

For example, when the model learns signals as well as noises in the training data but couldn't perform appropriately on new data upon which the model wasn't trained, the problem of overfitting takes place.

Various methods can be adopted, for avoiding overfitting of models on training data, such as cross-validation sampling, reducing number of features, pruning, regularization and many more.

DATA PROCESSING

FEATURE IMPORTANCE

Algorithms with the highest accuracy were taken, then the parameters that are most important for the final prediction results were identified. In other words, which parameters most influence the risk of diabetes.

04

RESULTS & DISCUSSION

Analysis of the obtained results

DATA PROCESSING RESULTS

	LR		KNN		SVM		NB		DT		RF		L2R	
	DD2019	**PID**	**DD2019**	**PID**	**DD2019**	**PID**	**DD2019**	**PID**	**DD2019**	**PID**	**DD2019**	**PID**	**DD2019**	**PID**
Accuracy	0.8899	0.7484	0.9541	0.7736	0.9587	0.7673	0.8257	0.7547	0.9541	0.7107	0.9541	0.7610	0.8807	0.7547
Error	0.1101	0.2516	0.0459	0.2264	0.0413	0.2327	0.1743	0.2453	0.0459	0.2893	0.0459	0.2390	0.1193	0.2453
Sensitivity	0.7937	0.4800	0.9206	0.4400	0.9365	0.4000	0.8095	0.5600	0.9365	0.5400	0.9206	0.4400	0.8095	0.4600
Specificity	0.9290	0.8716	0.9677	0.9266	0.9677	0.9358	0.8323	0.8440	0.9548	0.7890	0.9677	0.9083	0.9097	0.8899
Precision	0.8197	0.6316	0.9206	0.7333	0.9219	0.7407	0.6623	0.6222	0.8939	0.5400	0.9206	0.6875	0.7846	0.6571
10-fold CV	0.9006	0.7775	0.9649	0.7608	0.9312	0.7837	0.8562	0.7713	0.9634	0.7358	0.9527	0.7712	0.8945	0.7878

FEATURE IMPORTANCE

1 DD2019

- **K-Nearest Neighbor Classifier:** Gender, Stress sometimes, Regular medicine
- **Support Vector Machine:** Regular Medicine, Family Diabetes, Gender

2 PID

- **K-Nearest Neighbor Classifier:** Glucose, Insulin, Age
- **Support Vector Machine:** Insulin, BMI, Glucose

05

FURTHER WORK

Project Development Areas

THIS STUDY STILL REQUIRES FURTHER IMPROVEMENT AND EXPANSION OF THE FIELD OF STUDY. THE MAIN DIRECTION AT THE MOMENT IS THE USE OF OTHER MACHINE LEARNING ALGORITHMS, BUILDING MULTILAYER STRUCTURES AND CREATING THE NEURAL NETWORK. ALSO, A POSSIBLE AREA FOR FURTHER WORK IS THE COLLECTION OF OUR OWN DATABASE WITH AS MANY ATTRIBUTES AND PARTICIPANTS AS POSSIBLE.

Further work

06

CONCLUSION

Summary of the entire study

THE HIGHEST ACCURACIES ACHIEVED BY KNN AND SVM ON DD2019 (95.4%% AND 95.9%% RESPECTIVELY), WHILE ON PIDD KNN HAS 77.4% AND SVM HAS 76.7%. ON DD2019 FOR THE KNN THE MOST IMPORTANT ATTRIBUTES ARE GENDER, STRESS SOMETIMES, REGULAR MEDICINE, WHILE FOR SVM THESE ATTRIBUTES ARE REGULAR MEDICINE, FAMILY DIABETES, GENDER. ON PIDD GLUCOSE, INSULIN, AGE HAVE THE HIGHEST VALUES FOR THE KNN, WHILE FOR SVM THE MOST IMPORTANT ATTRIBUTES ARE INSULIN, BMI, GLUCOSE.

Conclusion

THE END OF PRESENTATION

THANK YOU!

Do you have any questions?

Feel free to ask questions

amanzhol.shungeyev@nu.edu.kz

REFERENCE LIST

- DeFronzo, R. A., Ferrannini, E., Groop, L., Henry, R. R., Herman, W. H., Holst, J. J., Hu, F. B., Kahn, C. R., Raz, I., Shulman, G. I., Simonson, D. C., Testa, M. A., & Weiss, R. (2015). Type 2 diabetes mellitus. Nature reviews. Disease primers, 1, 15019. https://doi.org/10.1038/nrdp.2015.19
- Normal blood sugar ranges and blood sugar ranges for adults and children with type 1 diabetes, type 2 diabetes and blood sugar ranges to determine people with diabetes. Diabetes.co.uk. (2019, January 15). Retrieved February 20, 2022, from https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html
- Shugart, C., Jackson, J., & Fields, K. B. (2010). Diabetes in sports. Sports health, 2(1), 29–38. https://doi.org/10.1177/1941738109347974

REFERENCE LIST

- Albright, A., Franz, M., Hornsby, G., Kriska, A., Marrero, D., Ullrich, I., & Verity, L. S. (2000). American College of Sports Medicine position stand. Exercise and type 2 diabetes. Medicine and science in sports and exercise, 32(7), 1345–1360. https://doi.org/10.1097/00005768-200007000-00024
- Agarwal, S. (2013). Data Mining: Data Mining Concepts and Techniques. 2013 International Conference on Machine Intelligence and Research Advancement, 203-207. doi: 10.1109/ICMIRA.2013.45
- Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, AA, Ogurtsova, K ., Shaw, JE, Bright, D., Williams, R., & IDF Diabetes Atlas Committee (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes research and clinical practice, 157, 107843. https://doi.org/10.1016/j.diabres.2019.107843

REFERENCE LIST

- UCI Machine Learning Repository: Data Set. Retrieved February 20, 2022, from https://archive.ics.uci.edu/ml/datasets/pima%2bindians%2bdiabetes
- Han, W., Shengqi, Y., Zhangqin, H., Jian, H., Xiaoyi ,W. (2018). Type 2 diabetes mellitus prediction model based on data mining, Informatics in Medicine Unlocked, volume 10, 100-107. https://doi.org/10.1016/j.imu.2017.12.006.
- Roshan, B., Ashish, K., Ritu, C., Harleen, K. (2019). Prediction and diagnosis of future diabetes risk: a machine learning approach, sci. 1, 1112. https://doi.org/10.1007/s42452-019-1117-9
- Farajollahi, B., Mehmannavaz, M., Mehrjoo, H., Moghbeli, F., & Sayadi, M. (2021). Diabetes Diagnosis Using Machine Learning. Frontiers in Health Informatics, 10(1), 65. doi:http://dx.doi.org/10.30699/fhi.v10i1.267

REFERENCE LIST

- Deepti S., Dilip S. (2018). Prediction of Diabetes using Classification Algorithms, Procedia Computer Science, volume 132, 1578-1585. https://doi.org/10.1016/j.procs.2018.05.122.
- Joshi, R. D., & Dhakal, C. K. (2021). Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. International journal of environmental research and public health, 18(14), 7346. https://doi.org/10.3390/ijerph18147346
- Sidong, W., Zhao, X., Chunyan, M. (2018). A comprehensive exploration to the machine learning techniques for diabetes identification, 291-295. DOI: 10.1109/WF-IoT.2018.8355130
- Aiswarya, I., Jeyalatha, S., Ronak, S. (2015). Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process, 5. 1-14. DOI: 10.5121/ijdkp.2015.5101

REFERENCE LIST

- Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. Frontiers in genetics, 9, 515. https://doi.org/10.3389/fgene.2018.00515
- Vijayan, V. & Anjali, C.. (2015). Prediction and diagnosis of diabetes mellitus — A machine learning approach, 122-127. DOI: 10.1109/RAICS.2015.7488400
- Neha, P. T., Shruti, G. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods, Procedia Computer Science, volume 167, 706-716. https://doi.org/10.1016/j.procs.2020.03.336.

REFERENCE LIST

- K. P. N. V. Satya, S., Karthik, J., Niharika, C., Srinivas, P., Ravinder, N., Prasad, C. (2021). Optimized Conversion of Categorical and Numerical Features in Machine Learning Models, 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 294-299, doi: 10.1109/I-SMAC52330.2021.9640967.
- Zhang, S., Wu, X., Zhu, M. (2010). Efficient missing data imputation for supervised learning, 9th IEEE International Conference on Cognitive Informatics (ICCI'10), 672-679, doi: 10.1109/COGINF.2010.5599826.
- Hasan, M., Alam, M., Das, D., Hossain, E., Hasan, M. (2020). Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers, in IEEE Access, 76516-76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

REFERENCE LIST

- Song, Yan-Yan & Lu, Ying. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry. 27. 130-5. 10.11919/j.issn.1002-0829.215044.
- Louppe, Gilles. (2014). Understanding Random Forests: From Theory to Practice. 10.13140/2.1.1570.5928.
- Demir-Kavuk, O., Kamada, M., Akutsu, T. Ernst-Walter K. (2011). Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features, BMC Bioinformatics 12, 412. https://doi.org/10.1186/1471-2105-12-412
- Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., Ridella, S. (2012). The 'K' in K-fold Cross Validation, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. http://www.i6doc.com/en/livre/?GCOI=28001100967420

REFERENCE LIST

- Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
- Karlijn, J., Viand, S., Johannes, B., Friedo, W., Carmine Zoccali, Kitty, J. (2009). Diagnostic methods I: sensitivity, specificity, and other measures of accuracy, Kidney International, 75 (12), 1257-1263. https://doi.org/10.1038/ki.2009.92.
- Hasan, K., Alam, M., Das, D., Hossain, E., Hasan, M. (2020). Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers, in IEEE Access, vol. 8, 76516-76531. doi: 10.1109/ACCESS.2020.2989857.

COMPANY PROFILE

ONLINE TECH SHOP

Here is where your presentation begins

CONTENTS OF THIS TEMPLATE

Here's what you'll find in this **Slidesgo** template:

1. A slide structure based on a company profile, which you can easily adapt to your needs. For more info on how to edit the template, please visit **Slidesgo School** or read our **FAQs**.
2. An assortment of graphic resources that are suitable for use in the presentation can be found in the **alternative resources slide**.
3. A **thanks slide**, which you must keep so that proper credits for our design are given.
4. A **resources slide**, where you'll find links to all the elements used in the template.
5. **Instructions for use**.
6. Final slides with:
    - The **fonts and colors** used in the template.
    - A **selection of illustrations.** You can also customize and animate them as you wish with the online editor. Visit **Storyset** to find more.
    - More **infographic resources**, whose size and color can be edited.
    - Sets of **customizable icons** of the following themes: general, business, avatar, creative process, education, help & support, medical, nature, performing arts, SEO & marketing, and teamwork.

You can delete this slide when you're done editing the presentation.

TABLE OF CONTENTS

01 ABOUT US

You can describe the topic of the section here

02 OUR SERVICES

You can describe the topic of the section here

03 OUR CLIENTS

You can describe the topic of the section here

04 EXPECTED PROJECTION

You can describe the topic of the section here

THE SLIDE TITLE GOES HERE!

Do you know what helps you make your point clear?
Lists like this one:

- They’re simple
- You can organize your ideas in a clear way
- You’ll never forget to buy milk!

And the most important thing: the audience won’t miss the point of your presentation

01 ABOUT US

You can enter a subtitle here if you need it

VIRTUAL 3D

COMPANY PROFILE

ABOUT US

You can give a brief description of the topic you want to talk about here. For example, if you want to talk about Mercury, you can say that it’s the smallest planet

A PICTURE ALWAYS REINFORCES THE CONCEPT

Images reveal large amounts of data, so remember: use an image instead of a long text. Your audience will appreciate it

OUR HISTORY

2000

Jupiter is a gas giant and the biggest planet in the Solar System

2005

Saturn is a gas giant. It’s composed mostly of hydrogen and helium

2010

Neptune is the farthest planet from the Sun and the fourth-largest

OUR HISTORY

2016

Despite being red, Mars is a very cold place full of iron oxide dust

2018

Mercury is the closest planet to the Sun and the smallest one

2020

Venus has a beautiful name and is the second planet from the Sun

OUR PHILOSOPHY

MISSION

Despite being red, Mars is a very cold place. It’s full of iron oxide dust

VISION

Jupiter is a gas giant and the biggest planet in the Solar System

VALUES

Mercury is the closest planet to the Sun and the smallest one

AWESOME WORDS


A PICTURE IS
WORTH A
THOUSAND WORDS

“This is a quote, words full of wisdom that someone important said and can make the reader get inspired.”

—SOMEONE FAMOUS

70,500

Big numbers catch your audience’s attention

LOCATIONS

SATURN

It’s a gas giant and has several rings

NEPTUNE

It’s the farthest planet from the Sun

10H 30M 10S

Mercury is the closest planet to the Sun and the smallest one in the Solar System

420,000,000

Venus has a beautiful name and is the second planet from the Sun. It’s terribly hot

500,000 KM

Earth is the third planet from the Sun and the only one that harbors life in the Solar System

OUR SERVICES

MERCURY

Mercury is the closest planet to the Sun

VENUS

Venus is the second planet from the Sun

MARS

Despite being red, Mars is a cold place

SATURN

Saturn is the only planet with rings

NEPTUNE

It’s the farthest planet from the Sun

JUPITER

It’s the biggest planet of all the Solar System

BEST SELLERS

CAMERA

Mercury is the closest planet to the Sun and the smallest one

1

TABLET

Venus has a beautiful name, but also high temperatures

2

MARS

Despite being red, Mars is a very cold planet full of iron oxide dust

3

OUR STRENGTHS

LOYALTY

Neptune is the farthest planet from the Sun

EFFICIENCY

Despite being red, Mars is a very cold place

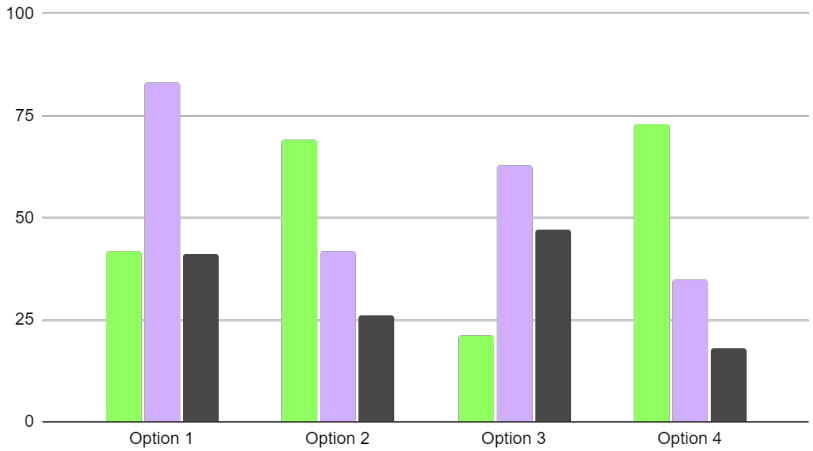
RELIABILITY

Mercury is the closest planet to the Sun

COMMIT

Jupiter is the biggest planet in the Solar System

OUR NUMBERS



Follow the link in the graph to modify its data and then paste the new one here. **For more info, click here**

MARS
Mars is actually cold

SATURN
Saturn is a gas giant

VENUS
Has a beautiful name

ONLINE SHOP
Web

100 UNITS FREE
Remaining

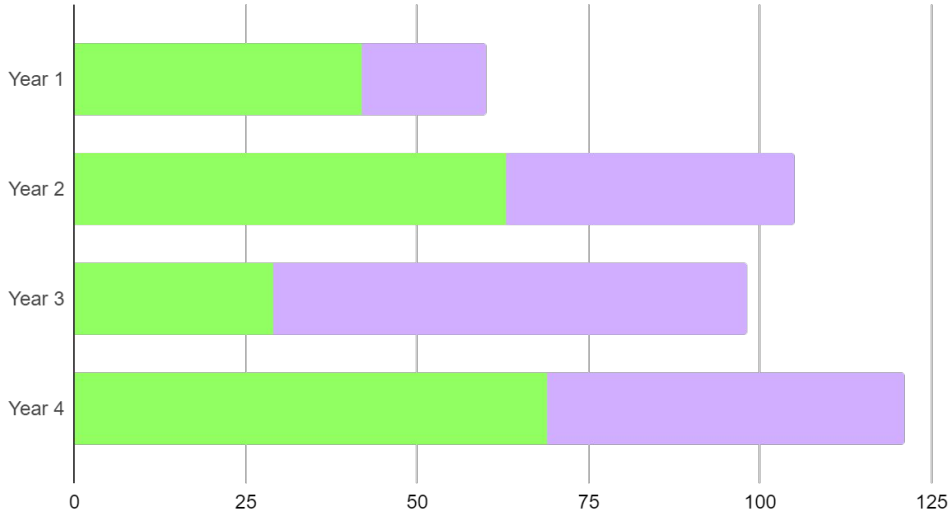
OUR GROWTH

$35,000

Saturn is a gas giant with rings

$6,000

Neptune is very far from the Sun



Follow the link in the graph to modify its data and then paste the new one here. **For more info, click here**

FUTURE PROJECTS

1

PROJECT 1

Jupiter is planet is the biggest planet in the entire Solar System

2

PROJECT 2

Venus has a beautiful name, but also high temperatures

3

PROJECT 3

Despite being red, Mars is a very cold planet full of iron oxide dust

4

PROJECT 4

Mercury is the closest planet to the Sun and the smallest one

CUSTOMER TESTIMONIALS

JOHN DOE

“Jupiter is a gas giant and the biggest planet in the Solar System”

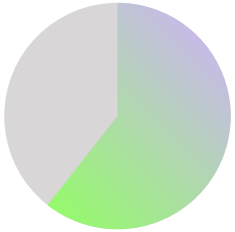
PETER PATS

“Mercury is the closest planet to the Sun and the smallest one”

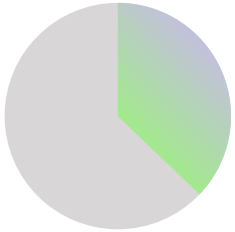
TARGET

AGE

GENDER



65% Male



35% Female

HOBBIES

AWARDS

MARS

Despite being red, Mars is a very cold place. It’s full of iron oxide dust

JUPITER

Jupiter is a gas giant and the biggest planet in the Solar System

MERCURY

Mercury is the closest planet to the Sun and the smallest one

OUR TEAM

JENNA DOE

You can talk a bit about this person here

JAMES WOLF

You can talk a bit about this person here

BANKING PRODUCTS

	BASIC	REGULAR	PRO
FREELANCERS			
WORKERS			
BIG COMPANIES			

DEVICES

You can replace the images on the screen with your own work. Just right-click on them and select “Replace image”

ONLINE TECH SHOP

THANKS!

Do you have any questions?

youremail@freepik.com

+91 620 421 838

yourcompany.com

ICON PACK: E-COMMERCE

ALTERNATIVE RESOURCES

PHOTOS

- An overhead view of cellphone with earphone; cd; radio and cassette
- Retro black camera arrangement

RESOURCES OF THIS TEMPLATE

Do you like the resources on this template? Get them for free at our other websites:

ICON PACK

- Icon Pack: E- Commerce | Outline

PHOTO

- Woman wearing virtual reality simulator
- Woman using her smartphone while at home
- Back view of woman at home using headphones and tablet
- Medium shot woman with headphones
- Smiley man at home on the couch using smartphone and headphones

Instructions for use

If you have a free account, in order to use this template, you must credit **Slidesgo** by keeping the **Thanks** slide. Please refer to the next slide to read the instructions for premium users.

As a Free user, you are allowed to:

- Modify this template.
- Use it for both personal and commercial projects.

You are not allowed to:

- Sublicense, sell or rent any of Slidesgo Content (or a modified version of Slidesgo Content).
- Distribute Slidesgo Content unless it has been expressly authorized by Slidesgo.
- Include Slidesgo Content in an online or offline database or file.
- Offer Slidesgo templates (or modified versions of Slidesgo templates) for download.
- Acquire the copyright of Slidesgo Content.

For more information about editing slides, please read our FAQs or visit Slidesgo School:

https://slidesgo.com/faqs and https://slidesgo.com/slidesgo-school

Instructions for use (premium users)

As a Premium user, you can use this template without attributing **Slidesgo** or keeping the "Thanks" slide.

You are allowed to:

- Modify this template.
- Use it for both personal and commercial purposes.
- Hide or delete the “Thanks” slide and the mention to Slidesgo in the credits.
- Share this template in an editable format with people who are not part of your team.

You are not allowed to:

- Sublicense, sell or rent this Slidesgo Template (or a modified version of this Slidesgo Template).
- Distribute this Slidesgo Template (or a modified version of this Slidesgo Template) or include it in a database or in any other product or service that offers downloadable images, icons or presentations that may be subject to distribution or resale.
- Use any of the elements that are part of this Slidesgo Template in an isolated and separated way from this Template.
- Register any of the elements that are part of this template as a trademark or logo, or register it as a work in an intellectual property registry or similar.

For more information about editing slides, please read our FAQs or visit Slidesgo School:

https://slidesgo.com/faqs and https://slidesgo.com/slidesgo-school

Fonts & colors used

This presentation has been made using the following fonts:

Bebas Neue
(https://fonts.google.com/specimen/Bebas+Neue)

Montserrat
(https://fonts.google.com/specimen/Montserrat)

Storyset

Create your Story with our illustrated concepts. Choose the style you like the most, edit its colors, pick the background and layers you want to show and bring them to life with the animator panel! It will boost your presentation. Check out How it works.

Pana

Amico

Bro

Rafiki

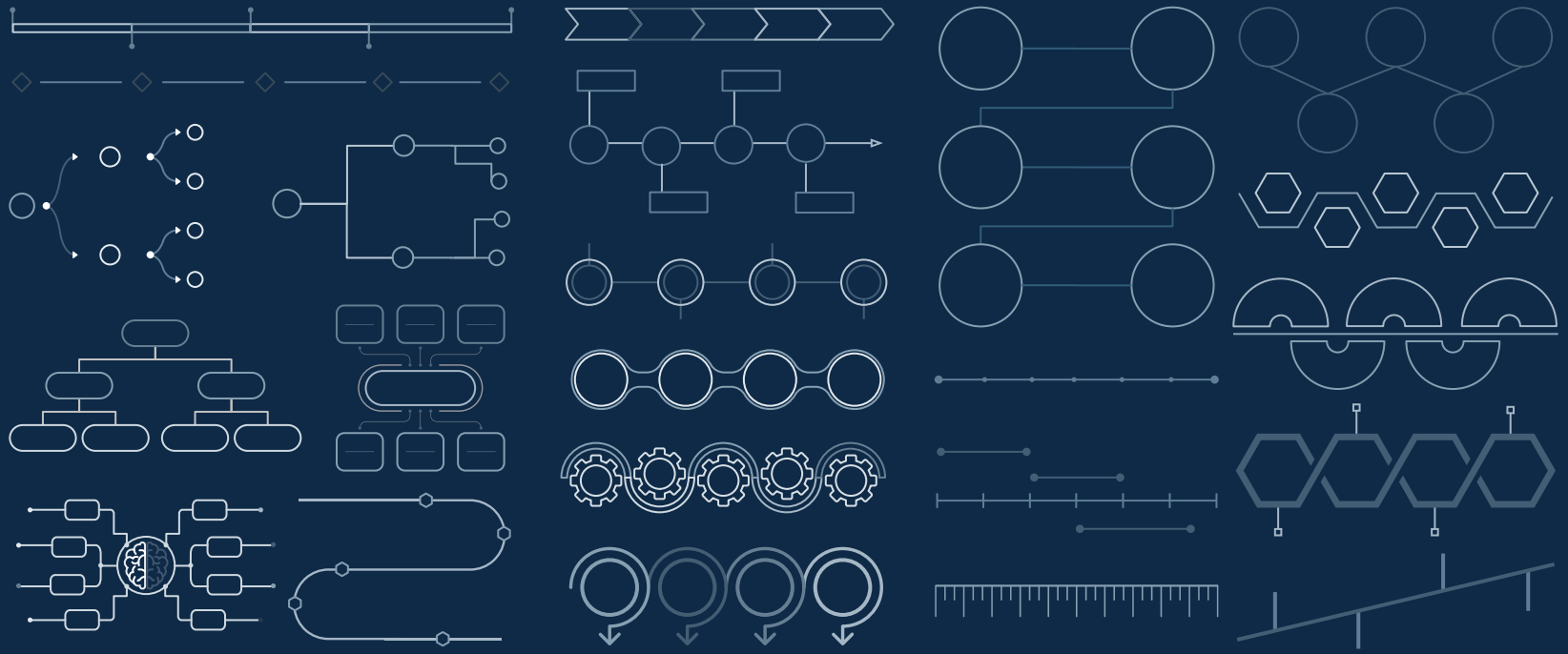
Cuate

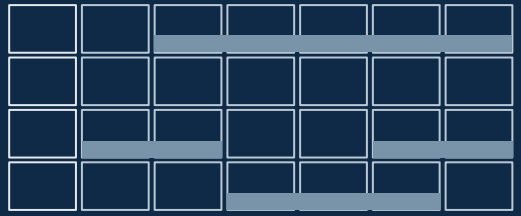
Use our editable graphic resources...

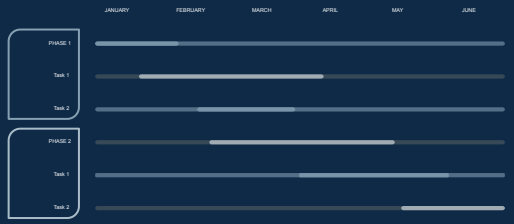
You can easily resize these resources without losing quality. To change the color, just ungroup the resource and click on the object you want to change. Then, click on the paint bucket and select the color you want. Group the resource again when you're done. You can also look for more infographics on Slidesgo.

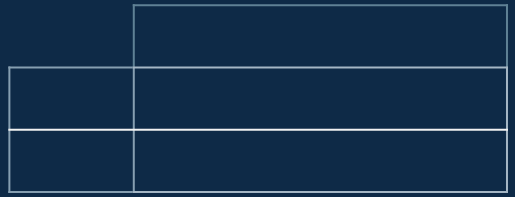


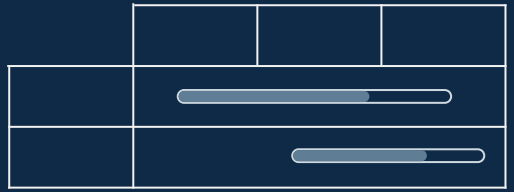


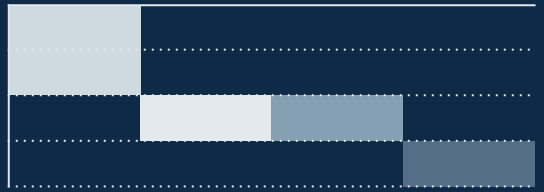



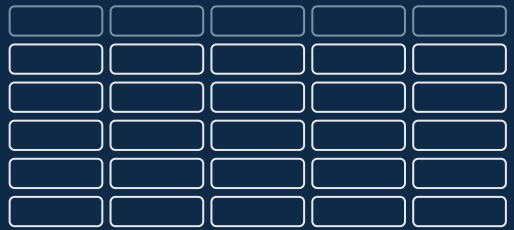


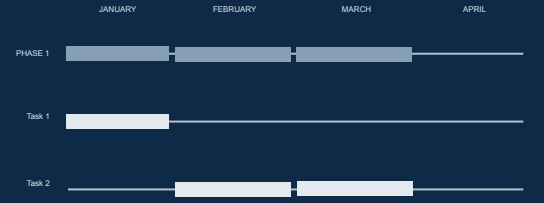

JANUARY
FEBRUARY
MARCH
APRIL
MAY
JUNE
PHASE 1
Task 1
Task 2
PHASE 2
Task 1
Task 2

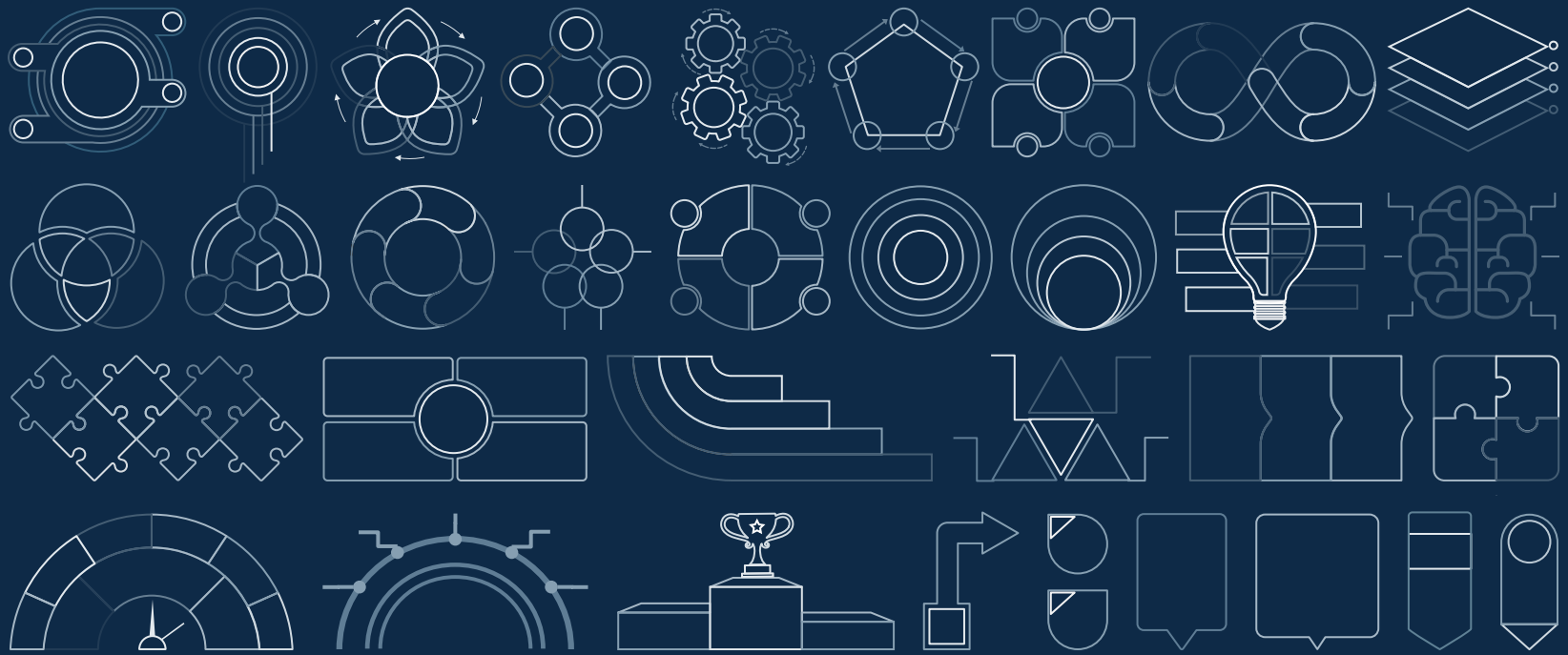


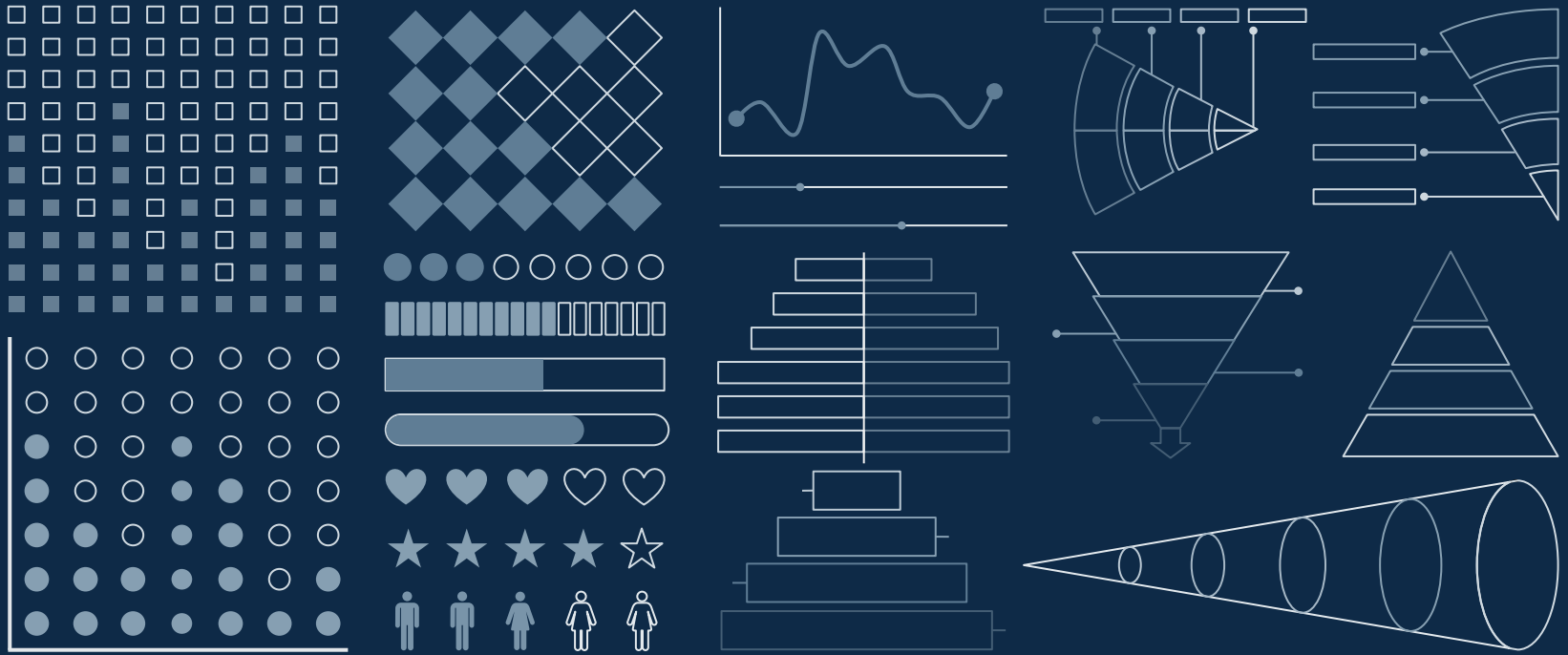







JANUARY
FEBRUARY
MARCH
APRIL
PHASE 1
Task 1
Task 2





...and our sets of editable icons

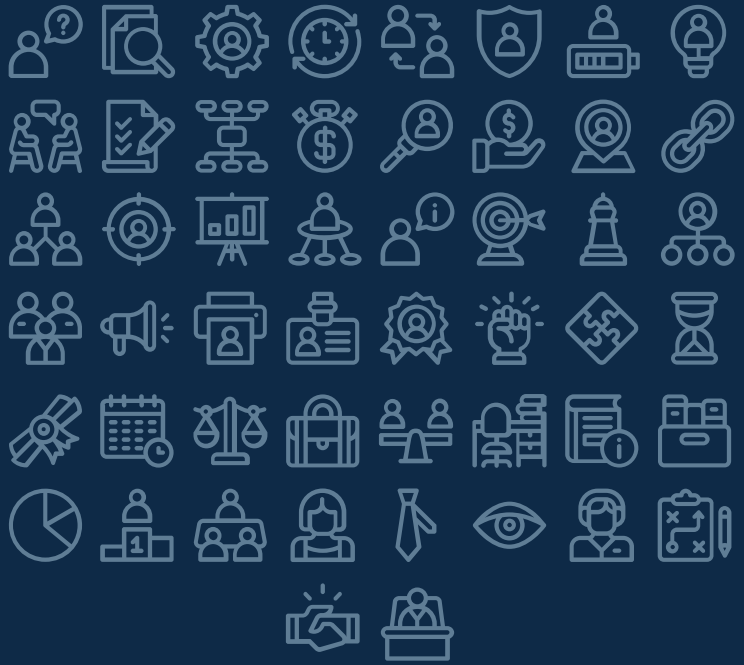
You can resize these icons without losing quality.
You can change the stroke and fill color; just select the icon and click on the paint bucket/pen.
In Google Slides, you can also use Flaticon's extension, allowing you to customize and add even more icons.

Educational Icons

Medical Icons

Business Icons

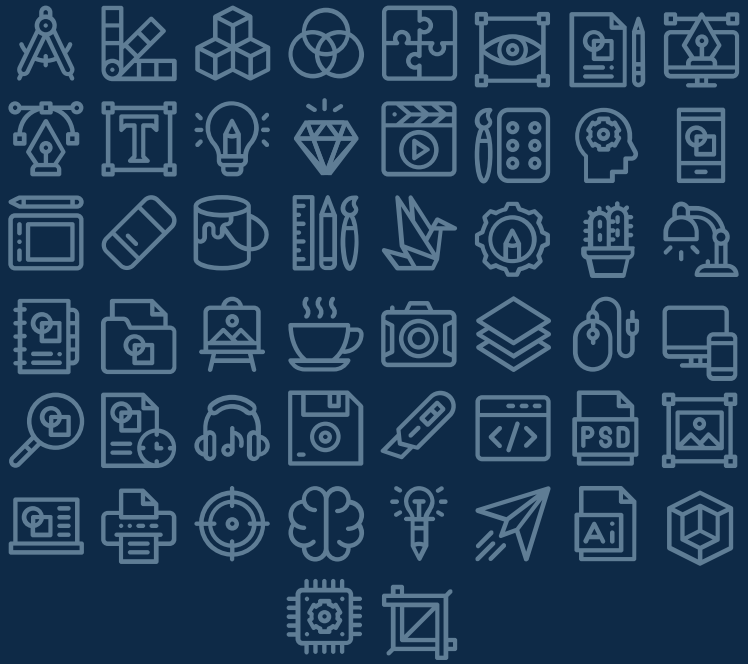
Teamwork Icons



Help & Support Icons

Avatar Icons

Creative Process Icons



Performing Arts Icons

Nature Icons

SEO & Marketing Icons


slidesgo