1. **PROBLEM ADDRESSED BY THE INVENTION:**

Diabetes is a chronic disease, which is becoming common and found in every third person near you. According to WHO, the Centres for Disease Control and Prevention (CDC) has indicated that as of 2018, 34.2 million Americans have diabetes and 88 million have prediabetes. Furthermore, the CDC estimates that 1 in 5 diabetics, and roughly 8 in 10 prediabetics are unaware of their risk. The CDC has collected a survey named, “Behavioral Risk Factor Surveillance System” in 2015 by 441,455 Americans on health-related risk behaviours, chronic health conditions, and the use of preventative services and generated 330 health indicators (features). In this project, we have considered to work with Health Indicator Diabetes Dataset to perform our analysis and early invention towards understanding of diabetes with different health features.

1. **OBJECTIVE OF THE INVENTION (Provide minimum two)**
2. **Objective 1:** To understand which health indicators are more relevant for a person to become diabetes in future and also how these indicators are correlated.
3. **Objective 2:** To develop a predictive machine learning (ML) model, which can prompt certain well-defined questions from users and able to predict the as the person is diabetes or non-diabetes.

**C. STATE OF THE ART/ RESEARCH GAP/NOVELTY:** Describe your invention fulfil the research gap?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sr. No. | Patent I’d | Abstract | Research Gap | Novelty |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

1. **DETAILED DESCRIPTION:**

In this project, we have built a predictive model which can help us identify the root cause on determine the factors which leads a person to become diabetic.

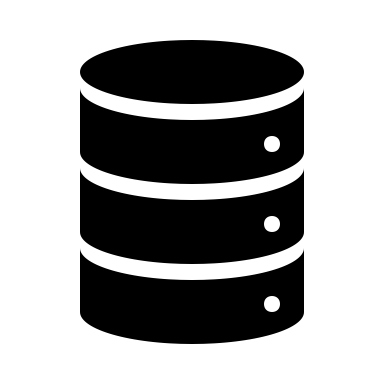
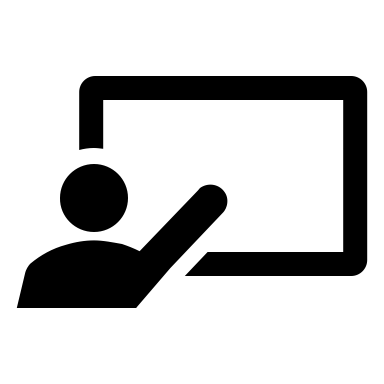
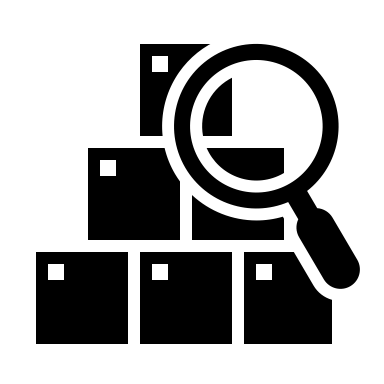
**Project Goals and Tools:**

* **Language** : Python3
* **Module**: Pandas, NumPy, Scikit-Learn, Matplotlib, Seaborn
* **Editor**: python-notebook, Google-Colab

**Research Methods :**

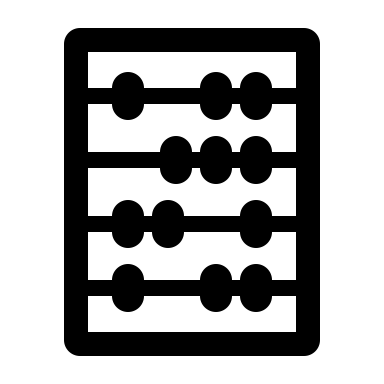
* Clean the dataset and perform the exploratory data analysis of the datafile, to understand the relevance of diabetes vs other health indicators
* Perform statistics analysis i.e. null hypothesis/ alternate hypothesis by applying methods like Chi-Square test to prove the feature relevance
* Build different model by removing outliers, build with all features, and build the model with important features
* Compare the different model by ROC curve, accuracy, precision and confusion Matrix
* Compare the model with respect to training time vs accuracy

Data Cleaning and Exploration

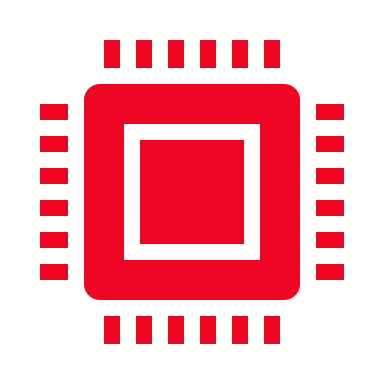
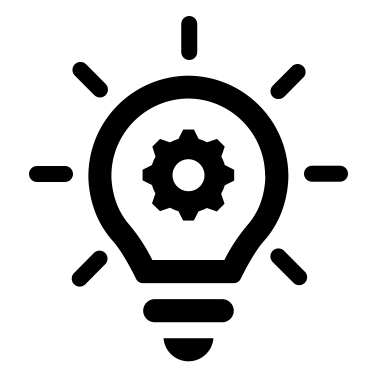
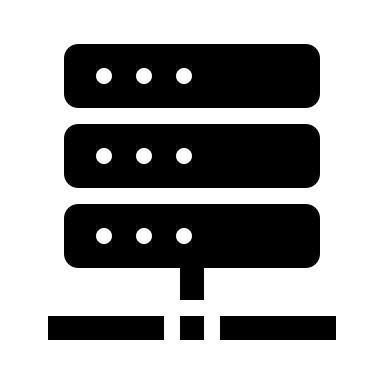


Diabetes Health Indicator Data

Evaluate



Build Model



User Input As Questions Answer

Predication

**Machine learning algorithm**

1. **RESULTS AND ADVANTAGES:**

* In data analysis, we found that few features are more important w.r.to Diabetes i.e. HighBP, BMI, GenHlth, MentHlth, PhyHlth, Age, Education and Income.
* The predictive machine learning model built by different classifiers i.e. Logistic Regression, Gaussian Naïve Bayes, Xgboost Classifier, Random Forest, and Decision Tree Classifiers.
* In term of training time and prediction accuracy, we measure that which classifier has the highest prediction accuracy and the which classifier achieve the minimum training time while compromising a little in accuracy.
* From this data analysis, we came to know that HighBP, Age, and BMI (obesity) is more importance health indicator which make a person diabetic.
* **Cost-effectiveness**: The model can be integrated into routine healthcare check-ups, or hospital for providing an affordable screening tool and early intervention to changes requires for healthy lifestyle.

1. **EXPANSION:**

To build the expansion of the project, we need to work on getting more data about the health i.e. blood pressure value, insulin value, and life style parameter (number of sleeping hours, eating hours). It will improve further the predictive power of the ML model. These data samples were collected only with Americans, there would be different impact of these indicators for other races i.e. African, Asians, Australian, etc. It could also impact the design of our predictive model for generalized diabetes predictive model.

1. **WORKING PROTOTYPE/ FORMULATION/ DESIGN/COMPOSITION:**

The current predictive model is functional and has been validated on publicly available diabetes health indicator datasets. However, further refinement with real-world clinical data would improve its robustness and generalizability. An estimated timeline of 6 months is required for full clinical testing and prototype validation.

1. **EXISTING DATA:**

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. It has been conducted every year since 1984. This original dataset contains responses from 441,455 individuals and has 330 features. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

In Health Indicator Diabetes Dataset has 3 data files, those details are given below:

* 1. **diabetes\_012\_health\_indicators\_BRFSS2015.csv:** It has 253,680 survey responses. The target variable Diabetes\_012 has 3 classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes. This dataset has 21 feature variables and is imbalanced.
  2. **diabetes\_binary\_5050split\_health\_indicators\_BRFSS2015.csv:** It has 70,692 survey responses. It has an equal 50-50 split of respondents with no diabetes and with either prediabetes or diabetes. The target variable Diabetes\_binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is balanced.
  3. **diabetes\_binary\_health\_indicators\_BRFSS2015.csv:** It has 253,680 survey responses. The target variable diabetes\_binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is not balanced.

It is the good dataset to observe and build predictive model as it has sufficient number of samples for training and testing. It has the following features i.e. Diabetes\_binary, HighBP, HighChol, CholCheck, BMI, Smokes, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education and Income.