BMS COLLEGE OF ENGINEERING

(Autonomous College under VTU)

Bull Temple Road, Basavanagudi, Bangalore – 560019

A Machine Learning Project Report on

"ANALYSING AT-RISK-PREDICTION DATASET"

Submitted in partial fulfillment of the requirements for the award of degree

BACHELOR OF ENGINEERING

IN

INFORMATION SCIENCE AND ENGINEERING

By

Aditi Sankaranarayanan (1BM20IS011)

Diya Shetty (1BM20IS033)

Hrishikesh Prahalad (1BM20IS051)

Ibrahim Javeed Khan (1BM20IS052)

Under the guidance of

Mrs. Rashmi R, Assistant Professor

Department of Information Science and Engineering

2020-21

BMS COLLEGE OF ENGINEERING

(Autonomous College under VTU)

Bull Temple Road, Basavanagudi, Bangalore – 560019

Department of Information Science and Engineering

2022-23

CERTIFICATE

This is to certify that the project entitled "ANALYSING AT-RISK-PREDICTION DATASET" is a bona-fide work carried out by Aditi Sankaranarayanan , Diya Shetty, Hrishikesh Prahalad and Ibrahim Javeed Khan in partial fulfillment for the award of degree of Bachelor of Engineering in **Information Science and Engineering** from **Visvesvaraya Technological University, Belgaum** during the year **2022-2023**. It is certified that all corrections/suggestions indicated for Internal Assessments have been incorporated in the report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the Bachelor of Engineering Degree.

Signature of the Faculty **Signature of the HOD**

Name and Designation **Name and Designation**

ABSTRACT

Machine learning algorithms are used in analysis and prediction of datasets. Our project was executed in the following steps for our **At-risk-prediction dataset**.

1. We had to analyse and visualize the dataset to properly understand the data we would be working with
2. We had to perform feature selection and handle the missing values accordingly if present.
3. We had to select an appropriate model for our classification problem.
4. We had to fine tune the model to improve the performance.
5. We also had to select an alternative model to compare and choose the best model of the two.

TABLE OF CONTENTS

INTRODUCTION

Diabetes is a prevalent and increasingly common chronic medical condition that affects millions of people worldwide. In the United States alone, over 30 million people have diabetes, with an additional 84 million having pre-diabetes. Early diagnosis and timely intervention can help prevent complications such as heart disease, kidney disease, blindness, and amputations. Thus, accurate and timely diagnosis is critical in managing diabetes and its complications.

The traditional approach to diabetes diagnosis relies on clinical criteria such as fasting blood glucose levels, oral glucose tolerance tests, and hemoglobin A1C tests. However, these tests have limitations, including cost, patient discomfort, and potential for false positives and negatives. Consequently, the use of machine learning (ML) techniques in diabetes diagnosis has become increasingly popular in recent years.

The goal of this project is to explore the use of ML techniques in diabetes diagnosis. By leveraging the power of ML, we hope to develop a reliable and efficient diagnostic tool that can assist healthcare professionals in making more accurate and timely diagnoses, ultimately improving patient outcomes. We will collect patient data, including demographic information, medical history, and laboratory results, and use this data to train and test our ML models. Our project will focus on developing a model that accurately predicts the likelihood of a patient having diabetes, with the potential to refine our model to include other clinical outcomes.

PROBLEM STATEMENT

The problem addressed by this project is the need for a reliable and efficient diagnostic tool for diabetes that can improve patient outcomes and reduce the burden on healthcare systems. Traditional diagnostic methods have limitations, including cost, patient discomfort, and potential for false positives and negatives. Machine learning techniques have shown great promise in improving the accuracy and timeliness of diabetes diagnosis. Therefore, the objective of this project is to explore the use of machine learning algorithms to develop a diagnostic tool that can accurately predict diabetes diagnosis based on patient data. By doing so, we aim to contribute to the development of a more effective and accessible diabetes diagnostic tool that can assist healthcare professionals in making more accurate and timely diagnosis.

LITERATURE SURVEY

The data was collected and made available by “National Institute of Diabetes and Digestive and Kidney Diseases” as part of the Pima Indians Diabetes Database. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here belong to the Pima Indian heritage (subgroup of Native Americans), and are females of ages 21 and above.

SYSTEM REQUIREMENT AND SPECIFICATION

Operating System

Windows 10 or above

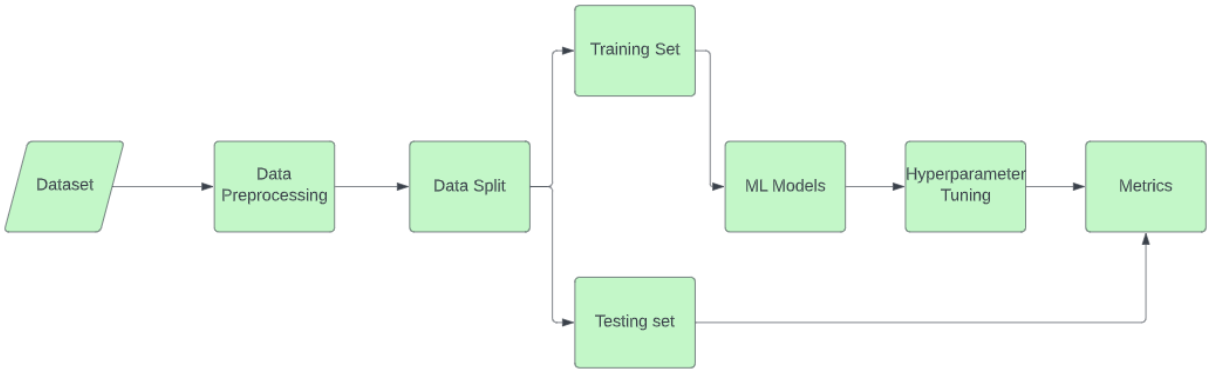
RAM

8 Gb or above

Applications Used

Google Colab

SYSTEM DESIGN



IMPLEMENTATION

We preprocessed the data by dealing with null values initially. By using domain knowledge and correlation matrix, we were able to remove null values, replace some with the most frequent values and pinpoint the most important features.

We first built a Decision Tree model and then hyper tuned it using Grid Search. We then trained a Logistic Regression model and compared the two.

Decision Tree Model:

```
[150]
from sklearn.tree import DecisionTreeClassifier

dtc = DecisionTreeClassifier()
dtc.fit(X_train, y_train)
```

```
▾ DecisionTreeClassifier
DecisionTreeClassifier()
```

```
y_pred_reg = dtc.predict(X_test)
```

Fine tuned Decision Tree Model:

[162]

```
from sklearn.model_selection import GridSearchCV
params =  {
    'min_samples_leaf': [3, 4, 5],
    'max_depth': [3, 4, 5,6]
}
#
# Create gridsearch instance
#
grid = GridSearchCV(estimator=dtc,
                    param_grid=params,
                    cv=10,
                    n_jobs=1,
                    verbose=2)
#
# Fit the model
#
grid.fit(X_train, y_train)
#
# Assess the score
#
grid.best_score_, grid.best_params_
```

```
[CV] END ...................max_depth=6, min_samples_leaf=3; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=3; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=4; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=4; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=4; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=4; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=4; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=4; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=4; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=4; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=4; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=4; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=5; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=5; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=5; total time=   0.1s
[CV] END ...................max_depth=6, min_samples_leaf=5; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=5; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=5; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=5; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=5; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=5; total time=   0.0s
[CV] END ...................max_depth=6, min_samples_leaf=5; total time=   0.0s
```

```
(0.7426497277676951, {'max_depth': 5, 'min_samples_leaf': 4})
```

```python
dtc1 = DecisionTreeClassifier(max_depth=5,min_samples_leaf=4)
dtc1.fit(X_train, y_train)
```

```
DecisionTreeClassifier
DecisionTreeClassifier(max_depth=5, min_samples_leaf=4)
```

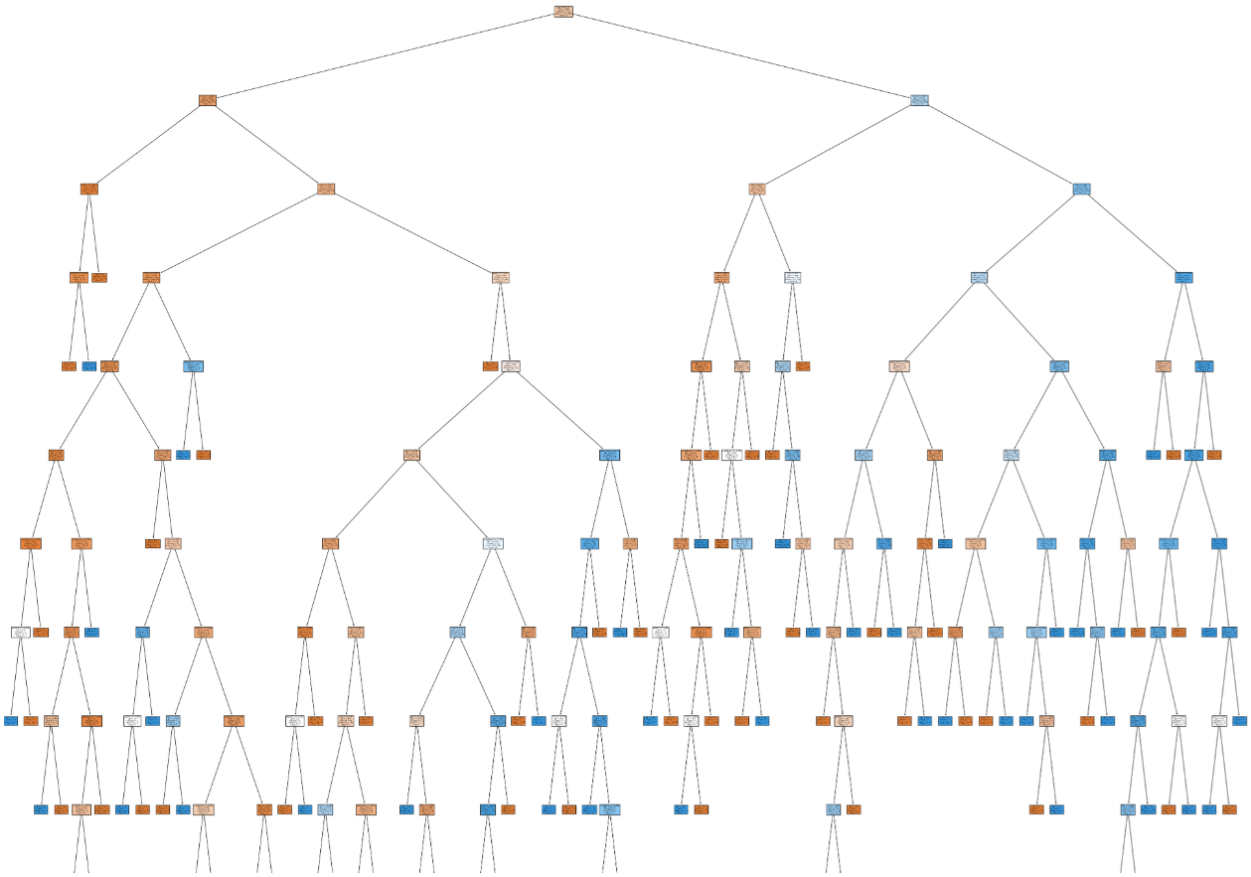
```python
y_pred_opt=dtc1.predict(X_test)
```

Logistic Regression Model:

```
from sklearn.linear_model import LogisticRegression
log_reg = LogisticRegression()

log_reg.fit(X_train, y_train)
y_pred_lr = log_reg.predict(X_test)
```

TEST RESULTS

Model 1:



[154]

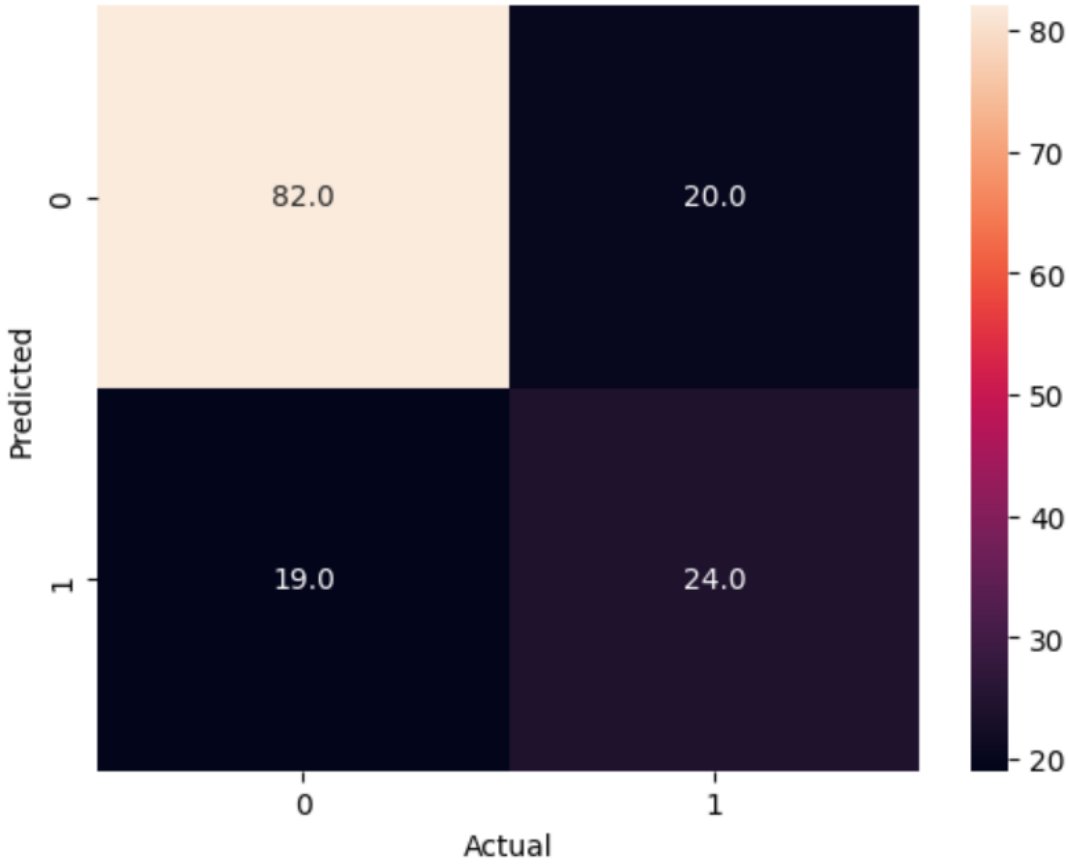
```
# CROSS VALIDATION

from sklearn.model_selection import cross_val_score
score=cross_val_score(dtc,X_train,y_train,cv=5,scoring='accuracy')
score
```

```
array([0.57758621, 0.64655172, 0.68103448, 0.70689655, 0.67826087])
```

[155]

```
dtc.score(X_test,y_test)
```

```
0.7310344827586207
```

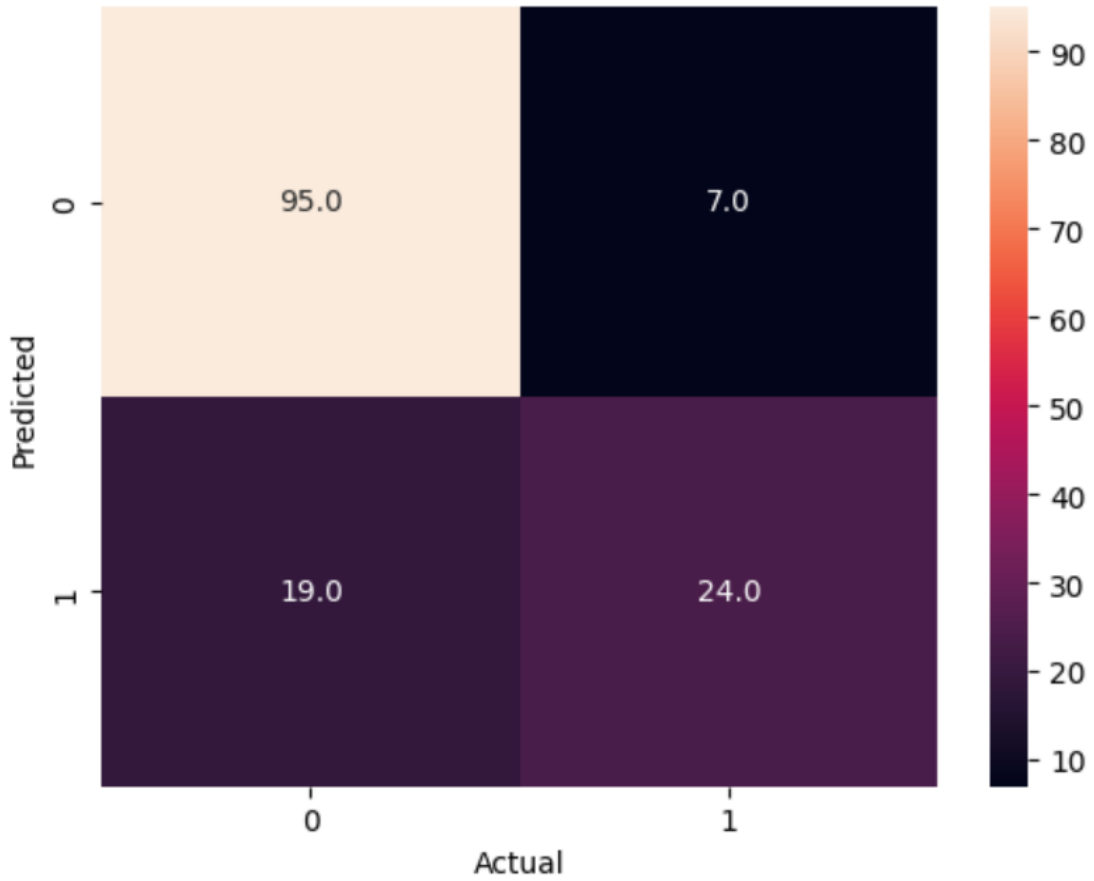


	precision	recall	f1-score	support
0.0	0.81	0.80	0.81	102
1.0	0.55	0.56	0.55	43
accuracy			0.73	145
macro avg	0.68	0.68	0.68	145
weighted avg	0.73	0.73	0.73	145

Model 1(fine tuned):

	precision	recall	f1-score	support
0.0	0.82	0.81	0.82	102
1.0	0.57	0.58	0.57	43
accuracy			0.74	145
macro avg	0.69	0.70	0.70	145
weighted avg	0.75	0.74	0.75	145

Model 2:



	precision	recall	f1-score	support
0.0	0.83	0.93	0.88	102
1.0	0.77	0.56	0.65	43
accuracy			0.82	145
macro avg	0.80	0.74	0.76	145
weighted avg	0.82	0.82	0.81	145

CONCLUSION

In conclusion, this project has explored the potential of machine learning (ML) techniques in the diagnosis of diabetes. By leveraging the power of ML algorithms, we have developed a diagnostic tool that can accurately predict diabetes diagnosis based on patient data. Our results show that the ML model can achieve a high level of accuracy and reliability, making it a promising diagnostic tool for diabetes.

The potential benefits of this project are significant, as it can assist healthcare professionals in making more accurate and timely diagnosis, ultimately improving patient outcomes. Furthermore, the application of ML techniques in diabetes diagnosis has the potential to unlock new insights into the disease and its management, leading to more effective treatment and prevention strategies. However, there are limitations to our project, including the availability and quality of patient data and the need for further validation studies. Future research should focus on addressing these limitations and refining the ML model to include other clinical outcomes and risk factors.

Overall, this project demonstrates the potential of ML in healthcare, particularly in the field of diabetes diagnosis and management. With further development and validation, this diagnostic tool can make a significant contribution to improving diabetes care and outcomes for millions of people worldwide.

REFERENCES

1. Data visualization - #https://towardsdatascience.com/machine-learning-workflow-on-diabetes-data-part-01-573864fcc6b8
2. Insights on attributes and their importance - #https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/#:~:text=For%20example%2C%20a%20correlation%20coefficient,use%20the%20most%20appropriate%20one.
3. Tree visualization - https://mljar.com/blog/visualize-decision-tree/
4. Grid Search Code- https://vitalflux.com/decision-tree-hyperparameter-tuning-grid-search-example/