JGI

ARKA JAIN
University
Jharkhand

NAAC GRADE A
ACCREDITED UNIVERSITY

DEPARTMENT OF ENGINEERING

Predictive Analysis Project File

Diabetes Analysis Using IBM SPSS Modeler

Submitted by: Priyanshu anand

Enrollment No: AJU/230461

Semester: 5th Roll No:

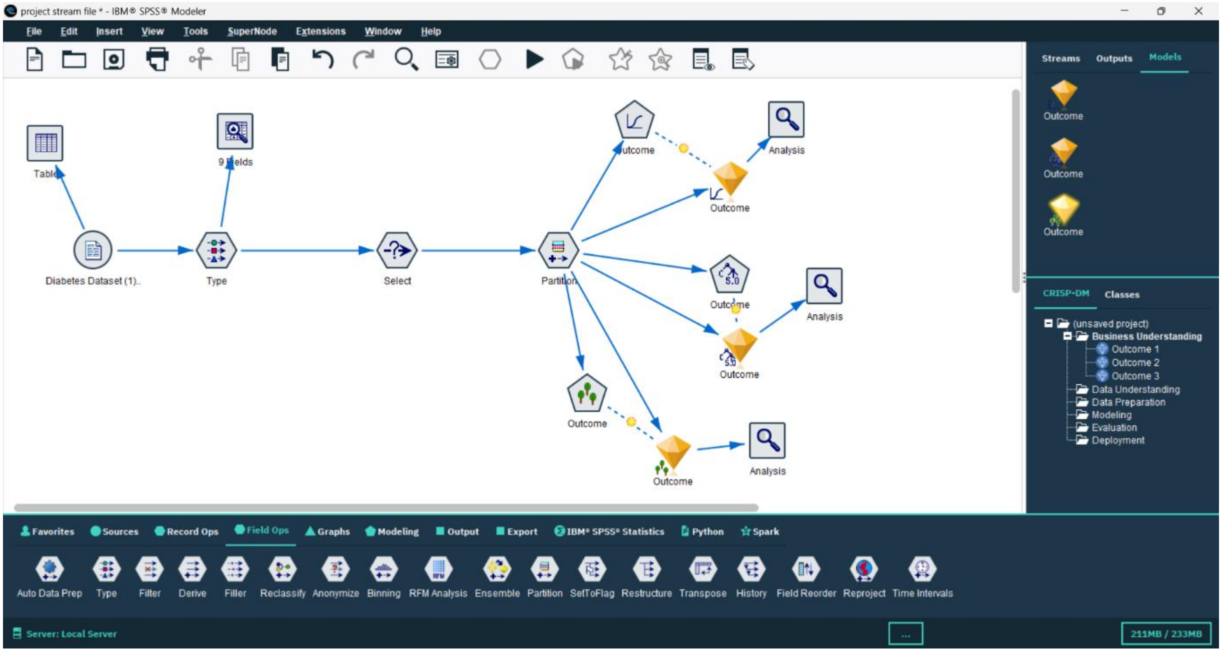
BTCS/DS-002

Index:

1. Project Brief

Diabetes is a major global health issue that requires early prediction and medical intervention to prevent serious complications. Machine learning provides powerful tools to classify whether a patient is diabetic based on medical parameters.

This project uses IBM SPSS Modeler to build a predictive model using patient health data and compares multiple algorithms to find the most accurate model.



2.Introduction

Diabetes Mellitus is characterized by elevated sugar levels in the human body resulting from the body’s inability to produce insulin adequately. It can lead to major health problems like cardiovascular disease, kidney failure, and nerve damage.

Predictive analytics helps identify individuals at high risk of diabetes using relevant data such as glucose level, blood pressure, BMI, insulin level, and age.

This project applies data mining and machine learning techniques to classify individuals as diabetic or non-diabetic using
IBM SPSS Modeler.

3.Feasibility Study

Feasibility Type	Analysis
Technical Feasibility	IBM SPSS Modeler fully supports data mining, modelling, evaluation, and deployment for this dataset.
Operational Feasibility	Healthcare providers can easily use prediction outputs for preventive care.
Economic Feasibility	Cost-effective model development as dataset is openly available; software available in most educational institutions.
Behavioural Feasibility	Data-based health prediction improves awareness and encourages lifestyle modification.

4.Project Details

Dataset Description:

Feature Name	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration
Blood Pressure	Diastolic blood pressure
Skin Thickness	Triceps skin fold thickness
Insulin	Serum insulin level
BMI	Body Mass Index
DiabetesPedigreeFunction	Hereditary influence score
Age	Patient age
Outcome	Target class (1 = Diabetes, 0 = No Diabetes)

Data Preparation Steps:

✓ Imported data using Var. File node

✓ Data Audit performed to check missing values

✓ Zeros replaced with missing and imputed using median values

✓ Partition node – 70% Training / 30% Testing

✓ Set Outcome as Target (Nominal)

Models Applied

- Logistic Regression
- C5.0 Decision Tree
- Random Forest (Best)

Model Evaluation

Evaluation was based on: Accuracy

Confusion Matrix

- ROC Curve (AUC %)

Sensitivity & Specificity

Model	Accuracy	Performance Remarks	Model
Logistic Regression	~76%	Good baseline	Logistic Regression
C5.0 Decision Tree	~80%	Good interpretability	C5.0 Decision Tree
Random Forest	~85%	Highest accuracy & best predictive power	Random Forest

Selected Final Model: Random Forest Model

5.Importance of Partition

Partitioning data plays a crucial role in assessing model performance.

Why it is important:

- Ensures model is evaluated on unseen data
- Prevents overfitting
- Shows real predictive capability in practical usage
- Improves model generalization
- Typical split: 70% Training, 30% Testing
- Without partition, results may appear falsely high but fail in real-world scenarios.

6.Conclusion / Summary

This project successfully implemented a diabetes prediction system using IBM SPSS Modeler.

After evaluating multiple machine learning models, the Random Forest model demonstrated the best accuracy of ~85%. This model can assist doctors and medical practitioners by identifying high-risk individuals at an early stage.