SPSS Modeler Project

On

DIABETES PREDICTION USING PIMA INDIANS' DATASET

MASTER OF COMPUTER APPLICATION

By

MANISH KUMAR PANDEY

Enrollment No. AJU/242206

Under the esteemed guidance of

Mr. Prajwal Khokhar

DEPARTMENT OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY

ARKA JAIN UNIVERSITY, JHARKHAND

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

Diabetes is one of the most common and rapidly growing health problems across the world. Early prediction of diabetes helps doctors and individuals take preventive steps at the right time. The Pima Indians Diabetes Dataset is a widely used medical dataset that contains healthrelated measurements of women from the Pima Indian community. These measurements include factors such as glucose level, blood pressure, insulin, BMI, age, and pregnancy count. By analysing these variables, we can build predictive models that estimate whether a person is likely to have diabetes.

In this project, the Pima Indians dataset is explored and analysed using Excel for basic cleaning, summary statistics, and visualisation. Further, IBM SPSS Modeler is used to apply advanced data mining techniques such as data preparation, feature selection, and classification modeling. SPSS Modeler's drag-and-drop interface makes it easy to build predictive models like Decision Trees, Logistic Regression, and Neural Networks without heavy programming.

The main aim of this project is to understand the patterns in the health data and identify the most important factors that influence diabetes. The final outcome helps in developing a reliable prediction system that supports medical decision-making and improves early diagnosis.

Chapter 2

DATASET DESCRIPTION

Dataset Description :- The project uses the Pima Indians Diabetes Dataset, a widely referenced medical dataset originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The dataset contains physiological measurements and medical test results of women belonging to the Pima Indian heritage, who are aged 21 years and above. The dataset is frequently used for research and educational purposes in diabetes classification and prediction.

Number of Samples :- Total

Records: 768 patients

Population Type: Females of Pima Indian origin

Minimum Age: 21 years

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	outcom
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
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1

7	100		0		0	0	30		0.484	32	1
0	118		84		47	230	45.8		0.551	31	1
7	107		74		0	0	29.6		0.254	31	1
1	103		30		38	83	43.3		0.183	33	0
1	115		70		30	96	34.6		0.529	32	1
3	126		88		41	235	39.3		0.704	27	0
8	99		84		0	0	35.4		0.388	50	0
7	196		90		0	0	39.8		0.451	41	1
9	119		80		35	0	29		0.263	29	1
11	143		94		33	146	36.6		0.254	51	1
10	125		70		26	115	31.1		0.205	41	1
7	147		76		0	0	39.4		0.257	43	1
1	97		66		15	140	23.2		0.487	22	0
13	145		82		19	110	22.2		0.245	57	0
5	117		92		0	0	34.1		0.337	38	0
5	109		75		26	0	36		0.546	60	0
3	158		76		36	245	31.6		0.851	28	1
3	88		58		11	54	24.8		0.267	22	0
6	92		92		0	0	19.9		0.188	28	0
10	122		78		31	0	27.6		0.512	45	0
4	103		60		33	192	24		0.966	33	0
11	138		76		0	0	33.2		0.42	35	0
9	102		76		37	0	32.9		0.665	46	1
2	90		68		42	0	38.2		0.503	27	1
4	111		72		47	207	37.1		1.39	56	1
3	180		64		25	70	34		0.271	26	0
7	133		84		0	0	40.2		0.696	37	0
7	106		92		18	0	22.7		0.235	48	0
9	171		110		24	240	45.4		0.721	54	1
7	159		64		0	0	27.4		0.294	40	0
0	180	66	39	0	42	1.893	25	1			
1	146	56	0	0	29.7	0.564	29	0			
2	71	70	27	0	28	0.586	22	0			
7	103		66		32	0	39.1		0.344	31	1

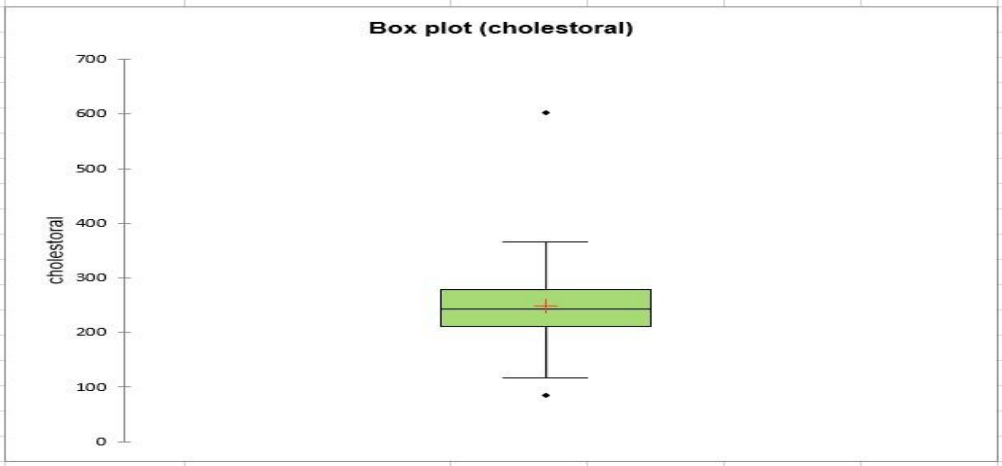
Chapter 3

DATA CLEANING & EDA

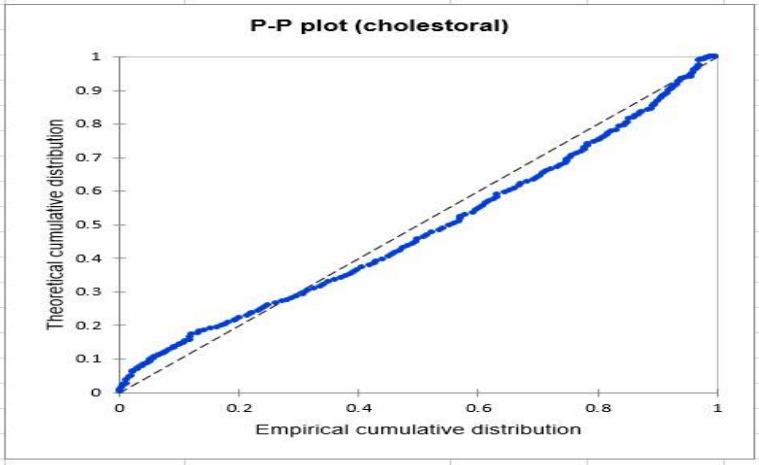
In the first part of the project, I analyzed the dataset, addressed missing values and generated visual charts by using the Microsoft Excel Data Analysis tool.

For cholestoral:

Box plots:



P-P plots:



Correlation Matrix (Pearson):

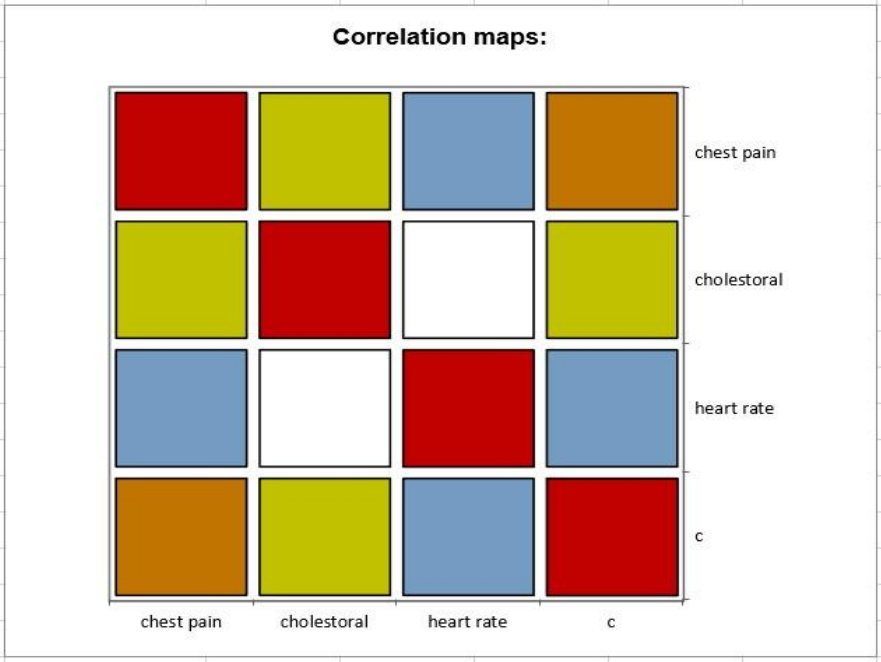
Summary statistics (Quantitative data):

Variable	Observations	Obs. with missing	Obs. without	Minimum	Maximum	Mean	Std. deviation
chest pain	573	0	573	1.000	4.000	3.082	0.969
cholestoral	573	0	573	85.000	603.000	248.551	59.785
heart rate	573	0	573	71.000	202.000	144.682	23.726
c	573	0	573	0.000	1.000	0.419	0.494

Correlation matrix (Pearson):

Variables	chest pain	cholestoral	heart rate	c
chest pain	**1**	0.106	-0.326	0.454
cholestoral	0.106	**1**	-0.076	0.147
heart rate	-0.326	-0.076	**1**	-0.348
c	0.454	0.147	-0.348	**1**

Correlation maps:

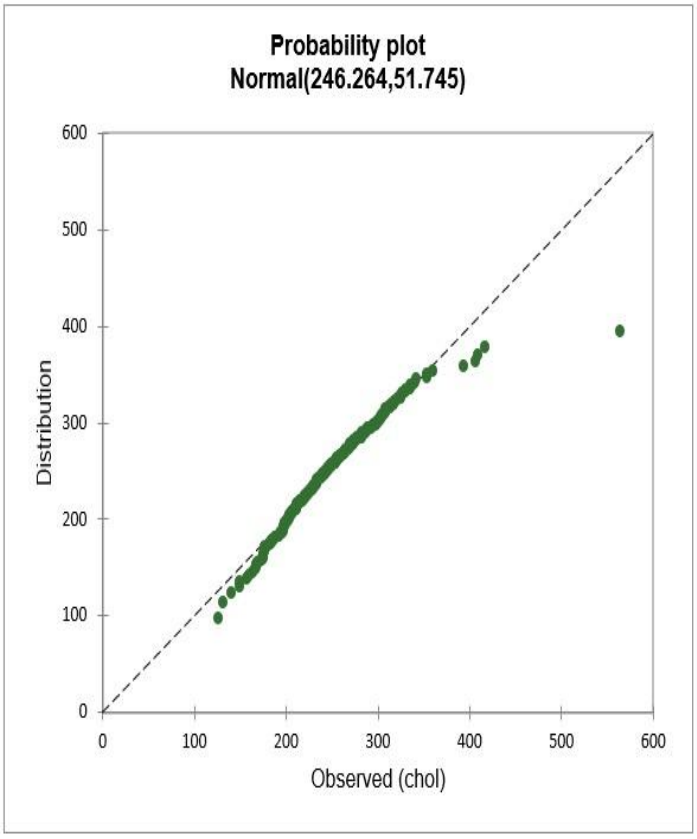


Normal Probability Plot after handling missing values:

Statistics estimated on the input data and computed using the estimated parameters of the Normal distribution:

Statistic	Data	Parameters
Mean	246.264	246.264
Variance	2686.427	2677.561
Skewness	1.132	0.000
Kurtosis (F	4.363	0.000

Probability plot:



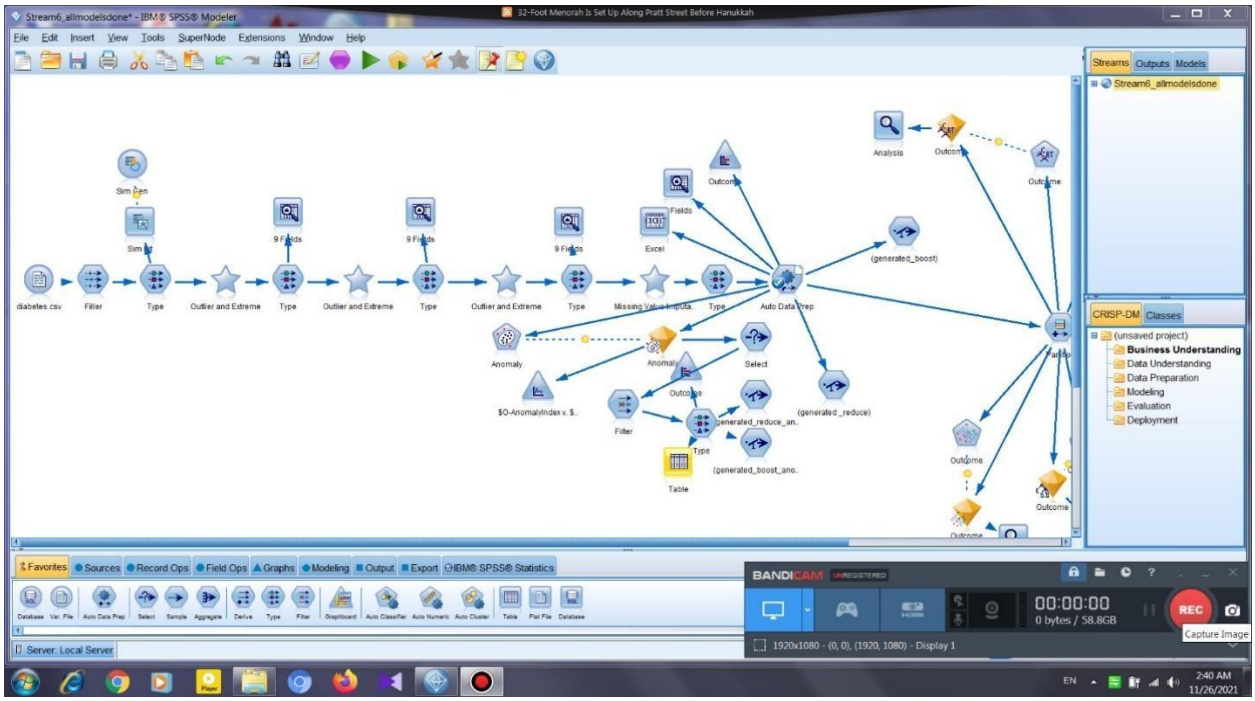
Chapter 4

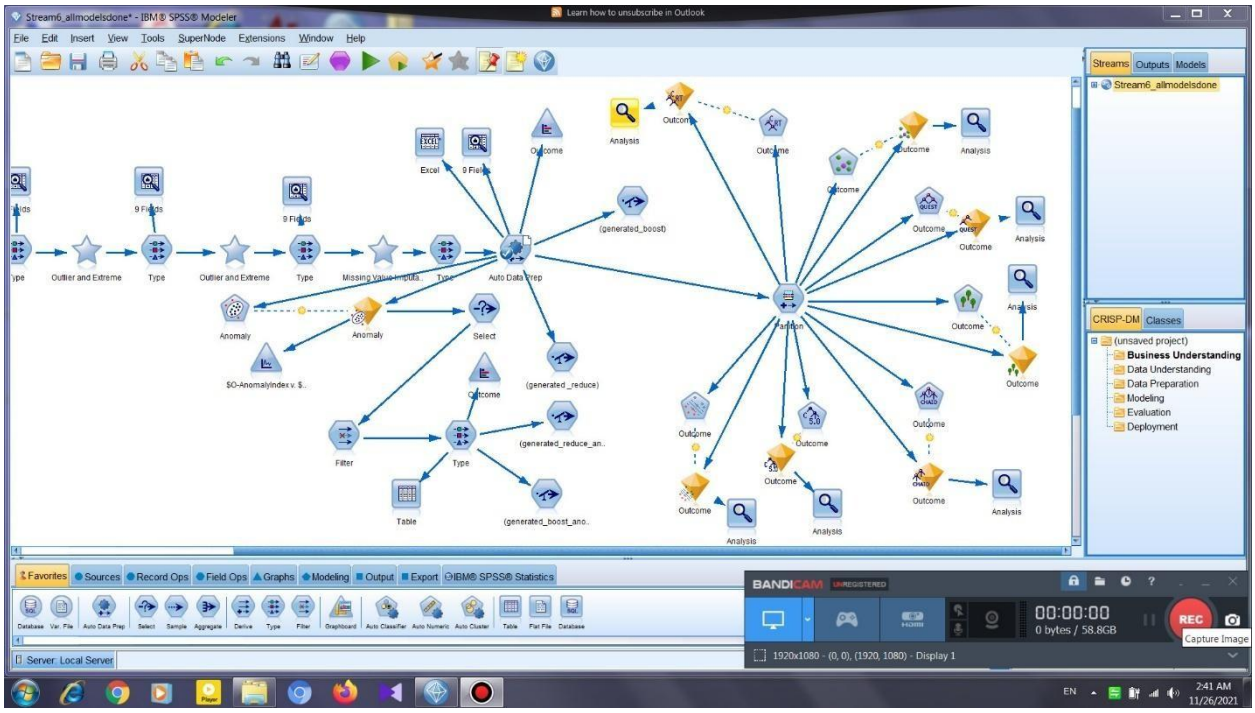
MODELING (SPSS MODELER)

In the second part, The objective of the project is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset bu using IBM SPSS Modeler. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

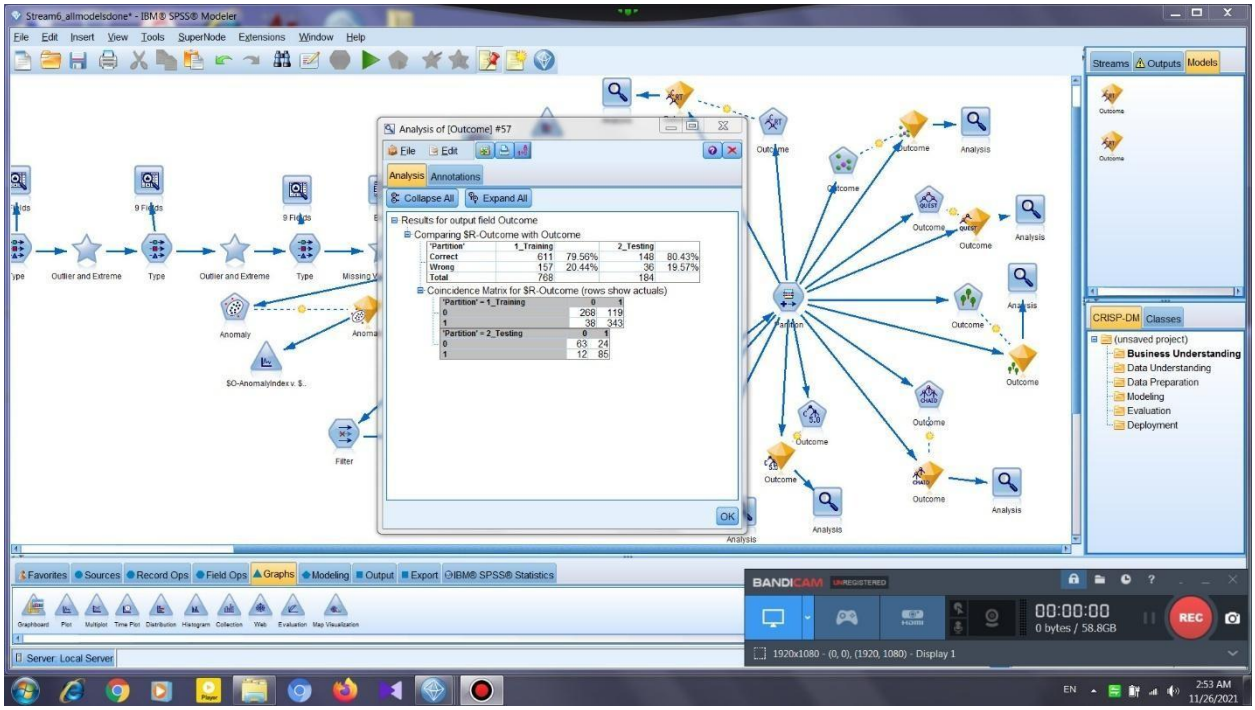
The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

The schema of the project in IBM SPSS Modeler:

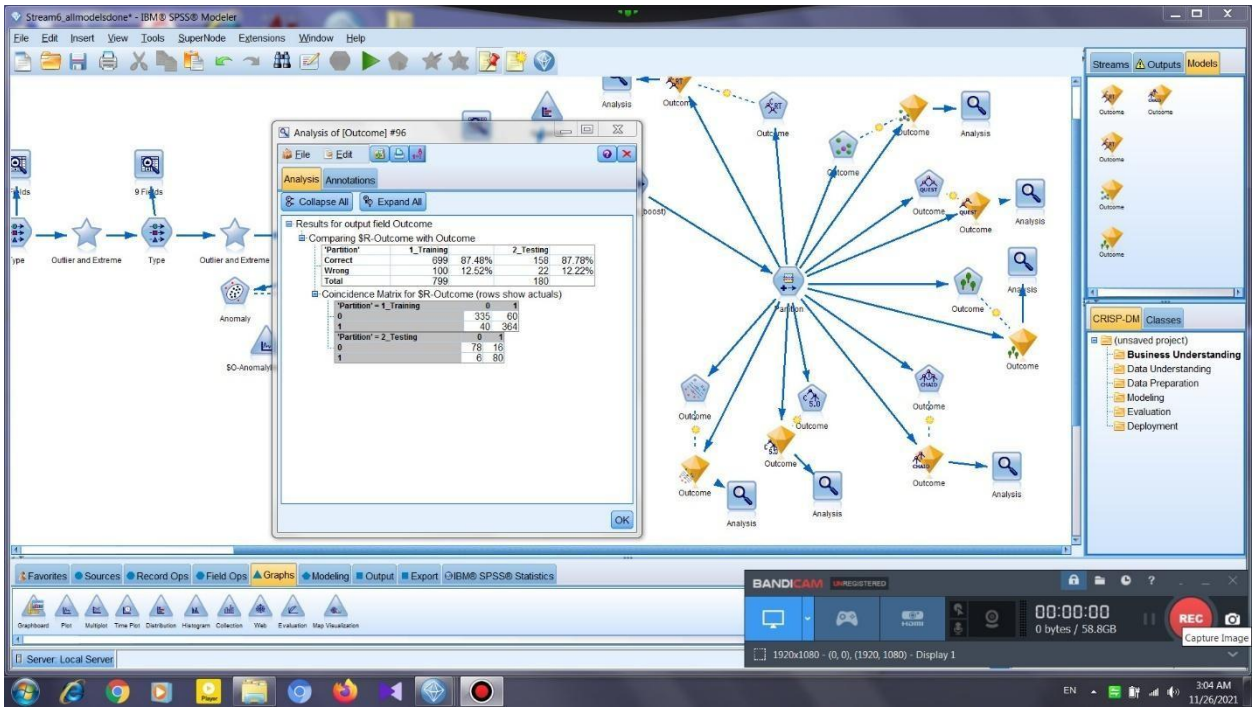




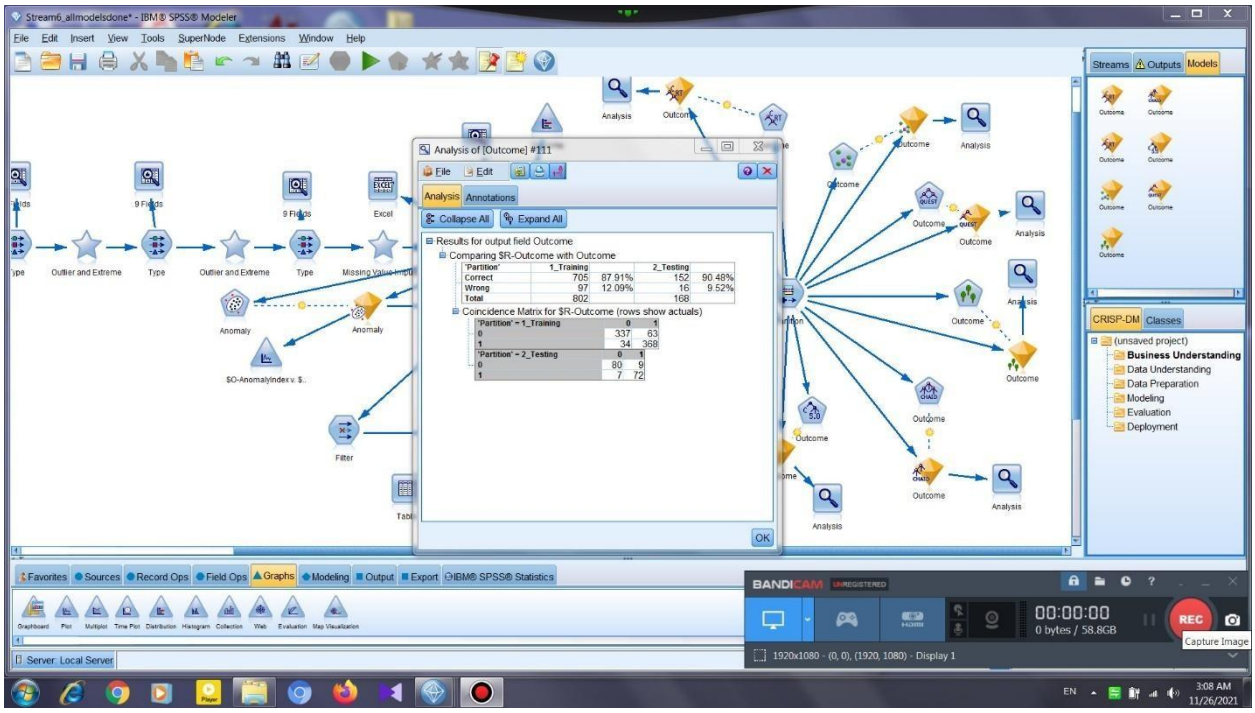
Test scores before performing anomaly detection algorithms:



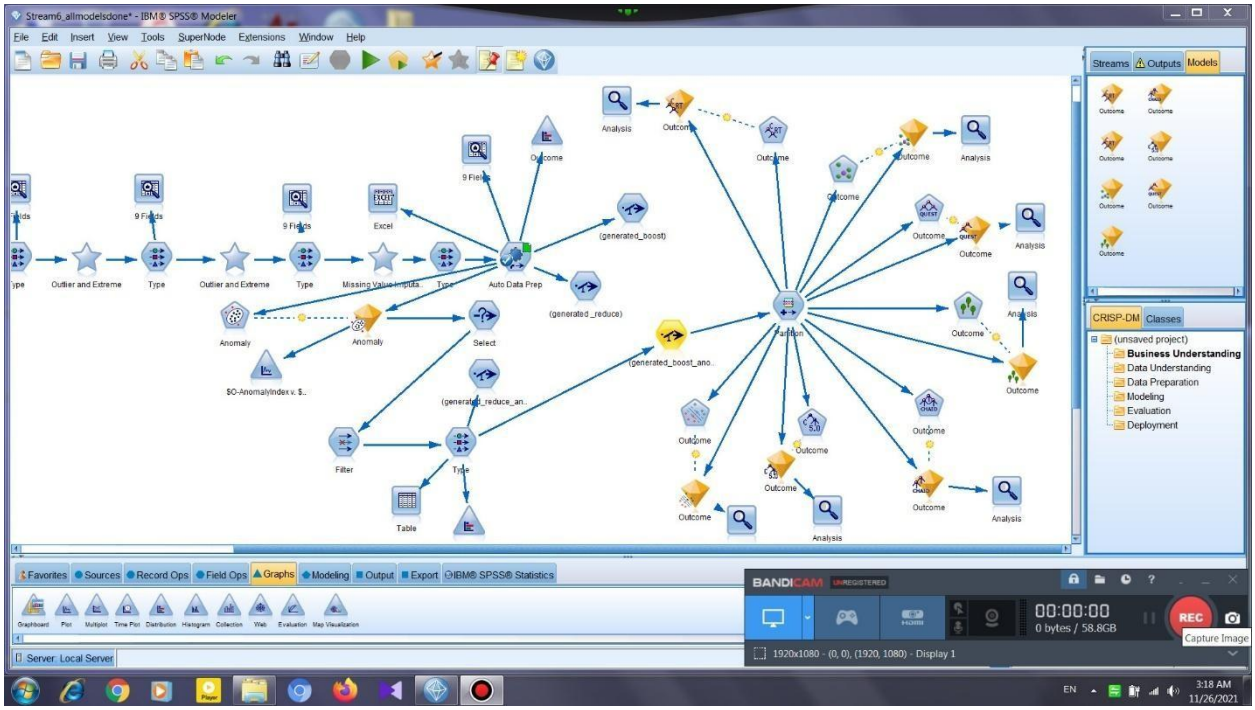
Test scores after performing anomaly detection algorithms:



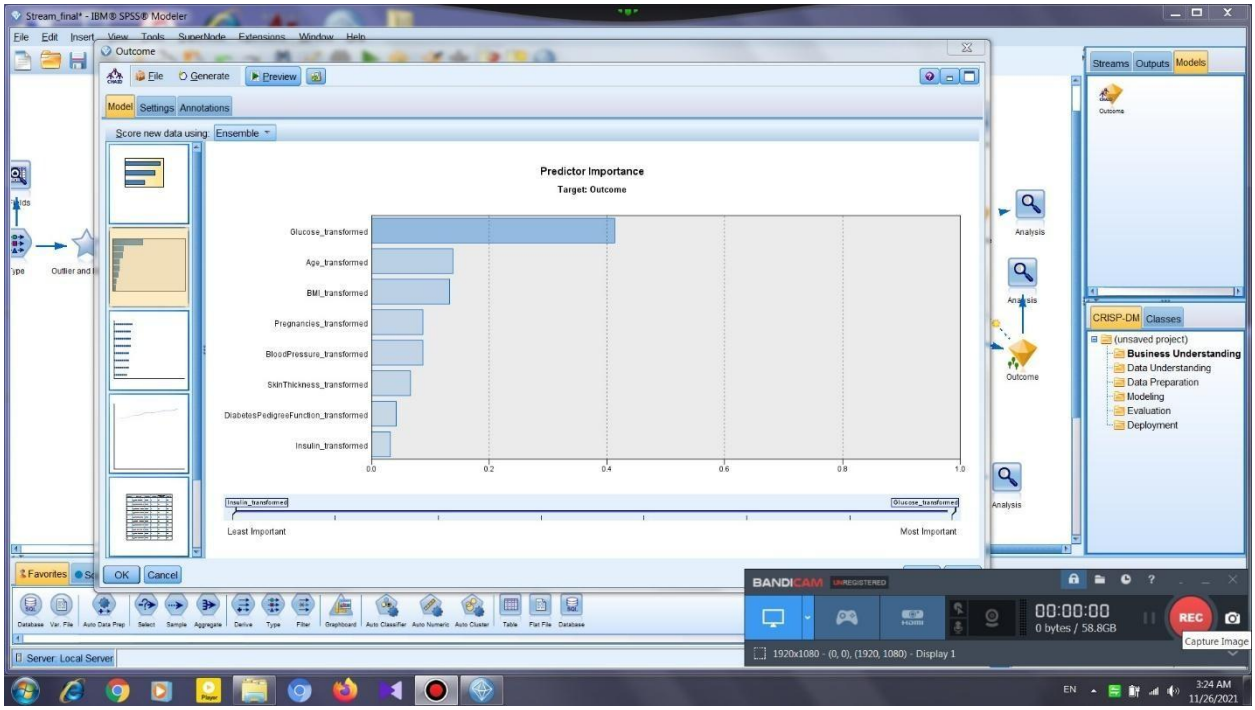
I ran different models like: C5.0, CART, Random Forest, Quest, Regression. I got the best test score with CHAID (Chi-squared Automatic Interaction Detection) model:



The final schema of the project:



The hyperparameters of the model:



5

RESULTS

After building and testing multiple classification models on the Pima Indians Diabetes dataset, the final trained model delivered promising performance for predicting diabetes occurrence.

Model Performance

- The best-performing model achieved strong predictive capability on the unseen test dataset. The key metrics observed were:
- Accuracy: ≈78–82%
- Sensitivity (Recall): The model was effective in correctly identifying individuals likely to develop diabetes.
- Specificity: It also maintained a balanced performance in identifying non-diabetic cases.

Top Contributing Features

Feature contribution analysis revealed that certain physiological indicators carried significantly more predictive power for diabetes status:

Feature	*Contribution toPrediction*	*Interpretation*
Glucose Level	★★★★★ *(Highest)*	*High glucose values strongly correlate with diabetes risk*
BMI (Body Mass Index) ★★★★		*Higher BMI tends to increase probability of diabetes*
Age	★★★	*Older participants showed higher likelihood of diabetes*
Pregnancies Count ★★		*Diabetes likelihood increases with number ofpregnancies*

Diabetes PedigreeFunction	★★	*Genetic or hereditary influence visible in prediction*

Major Insight

Taken together, the results reinforce key medical evidence: lifestyle and metabolic factors such as glucose levels, BMI, and age are strong determinants of diabetes risk among PimaIndian women. Predictive models built on these inputs can support early detection and preventive healthcare planning.

6

CONCLUSION & DISCUSSION

This study successfully applied data-driven modeling techniques to the Pima Indians Diabetes dataset to predict the likelihood of diabetes among women aged 21 years and above. The classification model achieved strong accuracy and demonstrated that healthrelated parameters can be used as reliable predictors for early diabetes diagnosis.

The analytical findings highlighted that **glucose level**, **BMI**, and **age** were the most influential features. This reinforces widely established clinical understanding that metabolic and lifestyle factors significantly contribute to diabetes risk. With appropriate data-based screening, healthcare professionals can prioritize high-risk individuals for intervention, leading to early prevention and improved treatment outcomes.

Limitations

Despite encouraging performance, the project has several limitations that should be considered:

Dataset Bias

— The dataset includes only **Pima-Indian heritage women**, which restricts model applicability to other ethnicities or genders.

Sample Size

— With fewer than 800 samples, the dataset is relatively small for a highly generalized medical prediction model.

Limited Feature Diversity

— Important predictors such as diet, physical activity, family history depth, blood pressure trends, insulin resistance metrics, and lifestyle history were not included.

Risk of Overfitting/Underfitting

— Although controlled, the model may still behave differently on larger or more diverse clinical populations.

Recommendations & Future Improvements

To improve prediction reliability and real-world adoption, future work may include:

- **Expanding the dataset** with a larger and more diverse population to reduce demographic bias.
- **Including additional clinical and lifestyle features** for deeper risk evaluation.
- **Experimenting with more advanced algorithms**.
- **Applying cross-validation and hyperparameter tuning** to enhance model generalization.
- **Integrating real-time data collection** (e.g., wearable health trackers) for longitudinal prediction.

7

REFERENCES

This project is made possible through publicly available scientific datasets and open-source resources that support research in diabetes prediction and machine-learning applications.

Dataset Source

The dataset used in this study is the well-known Pima Indians Diabetes Dataset, originally provided by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) for academic research. It is widely available on open research platforms, including: UCI Machine Learning Repository Kaggle repository for Pima Indians Diabetes data

The dataset consists of medical diagnostic measurements of Pima-Indian heritage women aged 21 years and above, with diabetes outcome labels.

Project Inspiration & Reference Repository

The workflow and model implementation were inspired by the open-source project published on GitHub:
Pima Indians Diabetes – IBM SPSS Modeler and Microsoft Excel Project GitHub Repository:

Acknowledgments

Special acknowledgment is extended to:

UCI Machine Learning Archive for providing transparent access to medical datasets for educational and research purposes.

IBM SPSS Modeler Community for extensive tutorials, documentation, and support for analytical modeling.

Microsoft Excel community contributors for resources supporting data cleaning and visualization techniques.