UNIVERSITY OF ASIA PACIFIC

Department of Computer Science and Engineering

Course Title: Artificial Intelligence and Expert Systems Lab
Course Code: CSE 404

Project No: 03

Submitted By:

Ripa Rani Biswas (20201001)

Md. Mominul Huque (20201049)

Anowar Hossen Farvez (20201059)

Submitted to:

Dr. Nasima Begum

Assistant Professor

Department of CSE,

University Of Asia Pacific.

Problem Description:

The task involves implementing multivariable linear regression using an open-source dataset. The objective is to compare the implementations done without using the SK-Learn library (raw code) and with SK-Learn. This comparison aims to understand the differences in implementation, performance, and ease of use between the two methods.

Tools and Languages:

---

1. Python
2. Google Colab

Dataset Selection:

Diabetes Dataset - https://drive.google.com/file/d/1m15FeURit9Rj1eYHcWzAZ_51ZsrMLdGO/view?usp=sharing

About Dataset

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on selecting these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

From the data set in the (.csv) File, We can find several variables, some of them are independent (several medical predictor variables) and only one target dependent variable (Outcome).

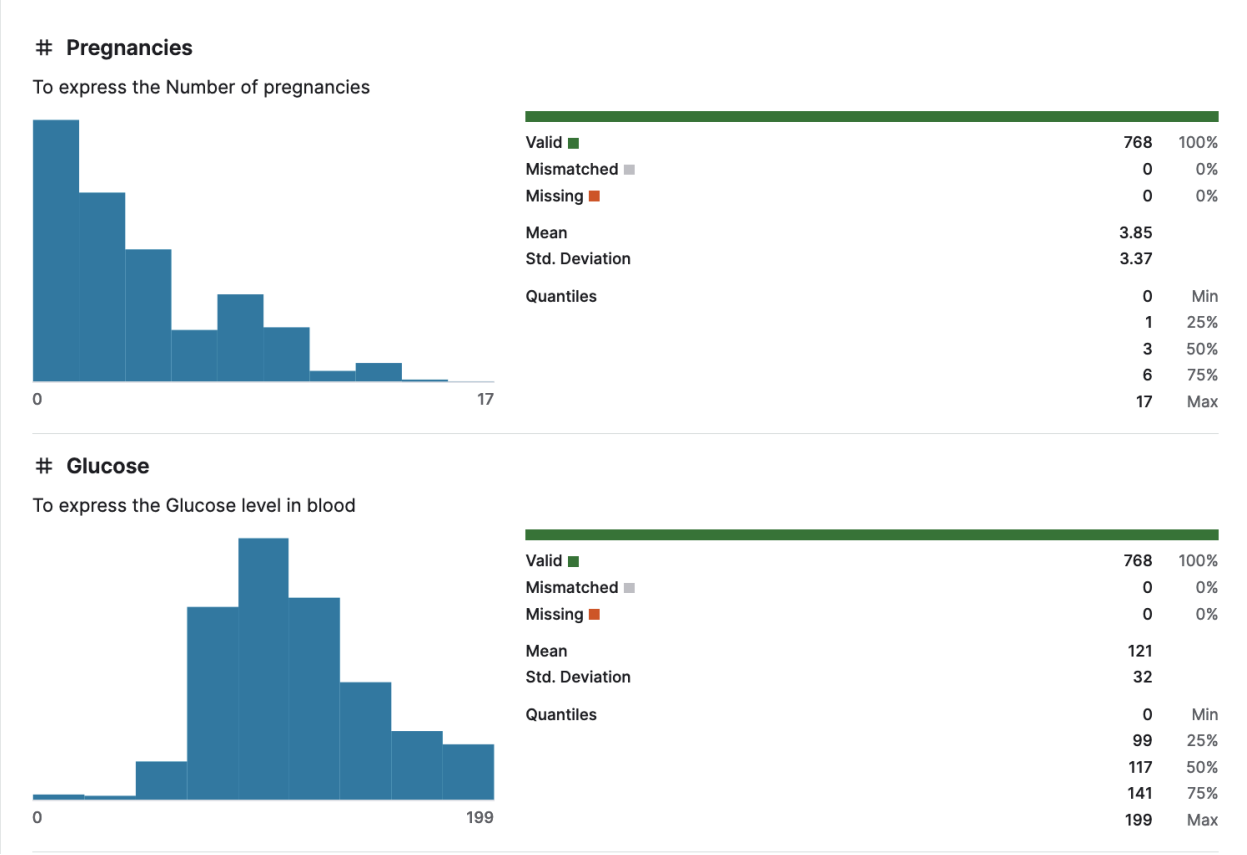
Information about dataset attributes -

---

1. **Pregnancies:** To express the Number of pregnancies
2. **Glucose:** To express the Glucose intake
3. **BloodPressure:** To express the Blood pressure measurement
4. **SkinThickness:** To express the thickness of the skin
5. **Insulin:** To express the Insulin level in the blood
6. **BMI:** To express the Body mass index
7. **DiabetesPedigreeFunction**: To express the Diabetes percentage
8. **Age:** To express the age
9. **Outcome:** To express the prediction of blood sugar

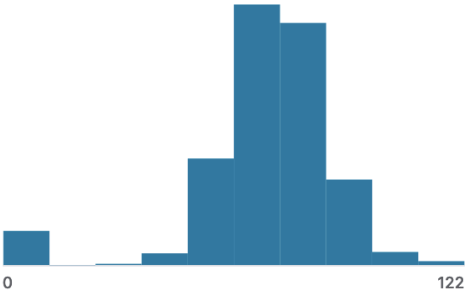
Dataset Preparation:

The dataset is well prepared with no null values and there is no need to scale the columns.



BloodPressure

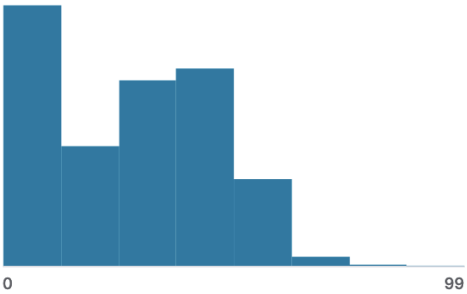
To express the Blood pressure measurement



Valid	768	100%
Mismatched	0	0%
Missing	0	0%
Mean	69.1	
Std. Deviation	19.3	
Quantiles	0	Min
	62	25%
	72	50%
	80	75%
	122	Max

SkinThickness

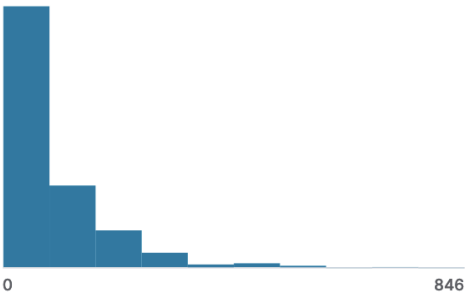
To express the thickness of the skin



Valid	768	100%
Mismatched	0	0%
Missing	0	0%
Mean	20.5	
Std. Deviation	15.9	
Quantiles	0	Min
	0	25%
	23	50%
	32	75%
	99	Max

Insulin

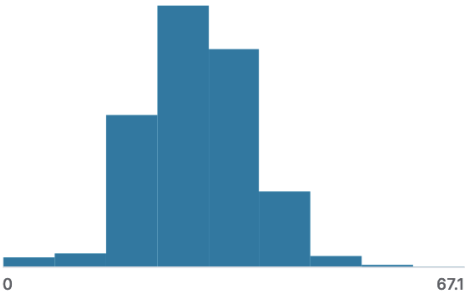
To express the Insulin level in blood



Valid	768	100%
Mismatched	0	0%
Missing	0	0%
Mean	79.8	
Std. Deviation	115	
Quantiles	0	Min
	0	25%
	32	50%
	128	75%
	846	Max

BMI

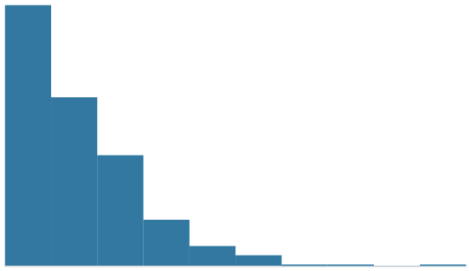
To express the Body mass index



Valid	768	100%
Mismatched	0	0%
Missing	0	0%
Mean	32	
Std. Deviation	7.88	
Quantiles	0	Min
	27.3	25%
	32	50%
	36.6	75%
	67.1	Max

DiabetesPedigreeFunction

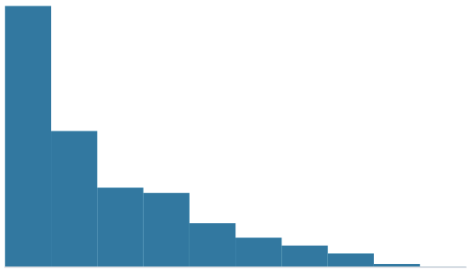
To express the Diabetes percentage



Valid	768	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.47	
Std. Deviation	0.33	
Quantiles	0.08	Min
	0.24	25%
	0.37	50%
	0.63	75%
	2.42	Max

Age

To express the age

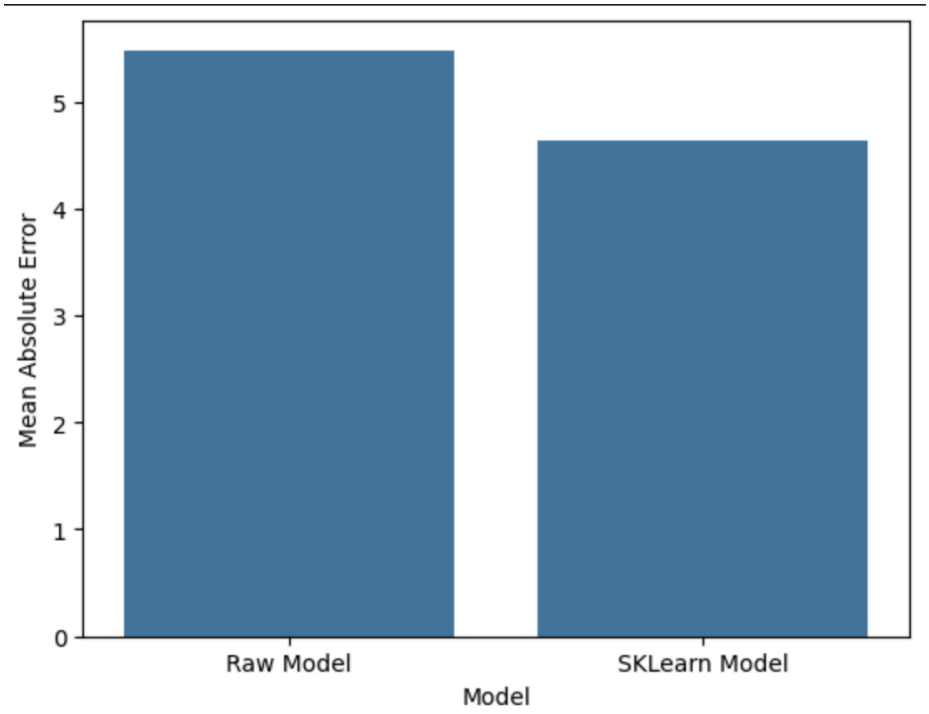


Valid	768	100%
Mismatched	0	0%
Missing	0	0%
Mean	33.2	
Std. Deviation	11.8	
Quantiles	21	Min
	24	25%
	29	50%
	41	75%
	81	Max

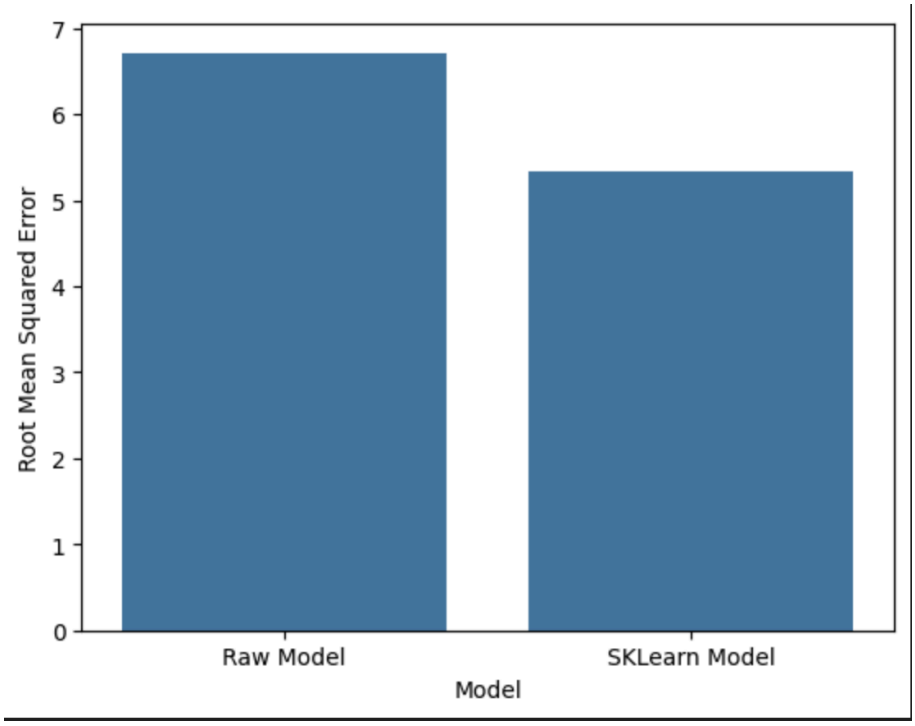
Comparison Table:

	Without SK-learn	With SK–learn
R^2 Score	-0.591	-0.002
Mean Absolute Error	5.479	4.640
Mean Squared Error	45.050	28.375
Root Mean Square Error	6.711	5.326

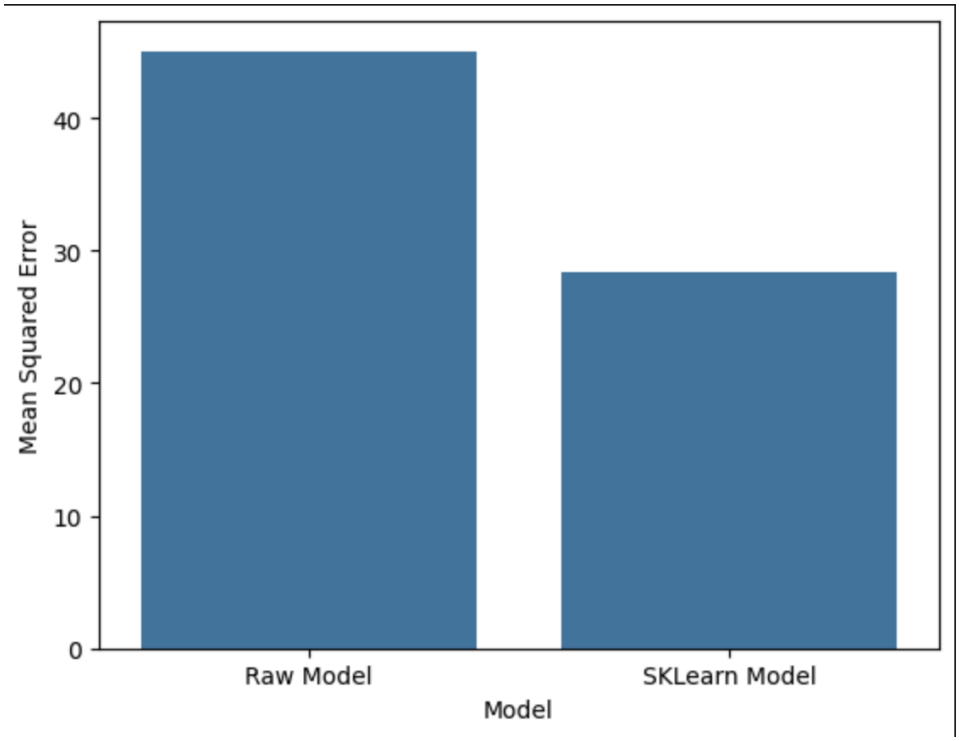
Mean Absolute Error Comparison:



Root Mean Squared Error Comparison:



Mean Squared Error Comparison:



We can see from the comparisons that the Scikit Learn Linear Regression Model implementation resulted in more accurate results. It's because Scikit Learn utilizes a highly optimized implementation, benefiting from efficient algorithms and data structures. This leads to faster training and prediction times compared to our implementations. It also offers regularization options like L1 (Lasso) and L2 (Ridge) regularization, and an automatic feature scaling feature that our implementation lacked.

Dataset and Model Suitability

We used the diabetes dataset to train the linear regression model, but both the hand-coded and Scikit learn models haven't performed well because the linear regression model isn't sufficient to figure out the interdependencies between the features.

Source Code: Project03-Regression.ipynb

Conclusion

Multivariable regression is an extension of simple linear regression. In this project, we learned a brief knowledge of linear regression after successful implementation. Multivariable linear regression models are useful in helping an enterprise consider the impact of multiple independent predictors and variables on a dependent variable and can be beneficial for forecasting and predicting results. The outcome knowledge of this project will help us a lot in real-life problem predicting and solving in various fields.