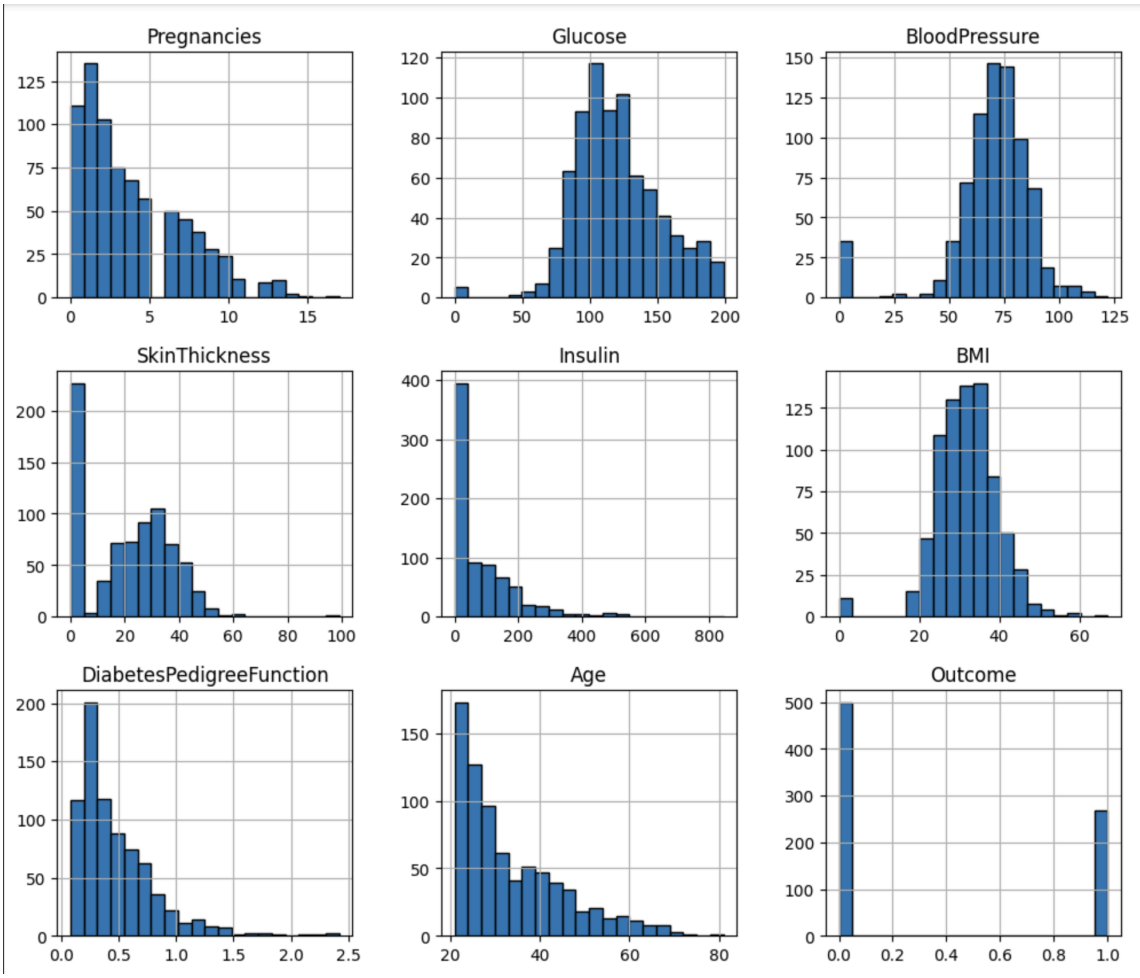


The key insights of above EDA analysis of dataset:

The data is **negatively skewed** (i.e there are more negative cases - non diabetes (0) than that of positive- has diabetes (1)

Pregnancies: Most patients had **0-2 pregnancies**, fewer cases with higher counts.

Glucose Levels: Slightly **right-skewed**, indicating some patients have very high glucose levels.

Blood Pressure: Nearly **normal distribution**, centered around 70-80.

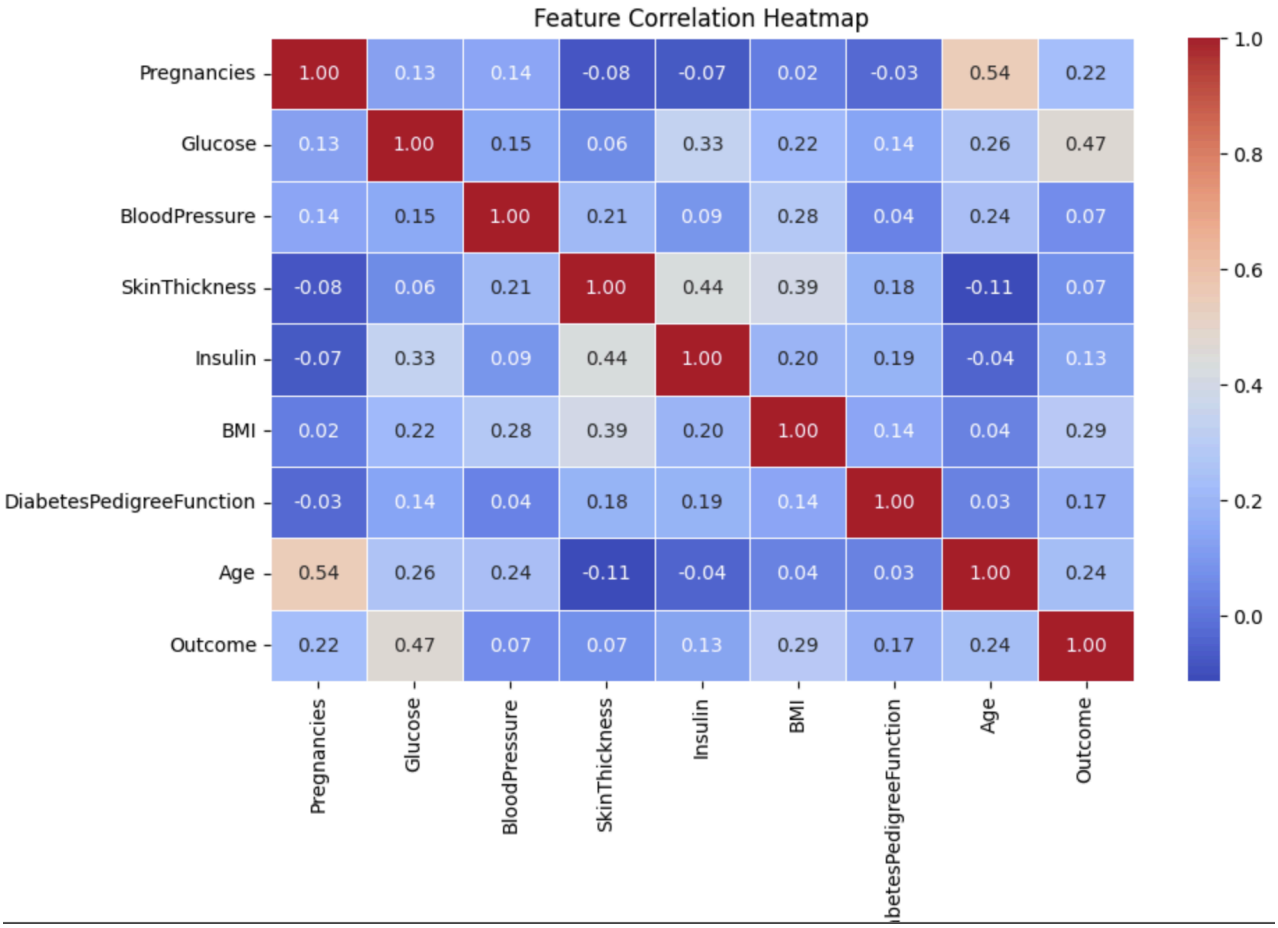
BMI: Most values around 25-35, with some extreme values.

Age: Majority of patients are in the **20-40 age range**, with fewer older patients.

Diabetes Pedigree Function (Genetic Risk): Right-skewed, meaning most patients have a lower genetic risk score.

Missing/Incorrect Values: Insulin & Skin Thickness have many zero values → **Possible missing data** which can be fixed using **mean/median imputation** .

Outliers: Glucose, Insulin, BMI, Age have extreme values. Applying **log transformation** can fix this.



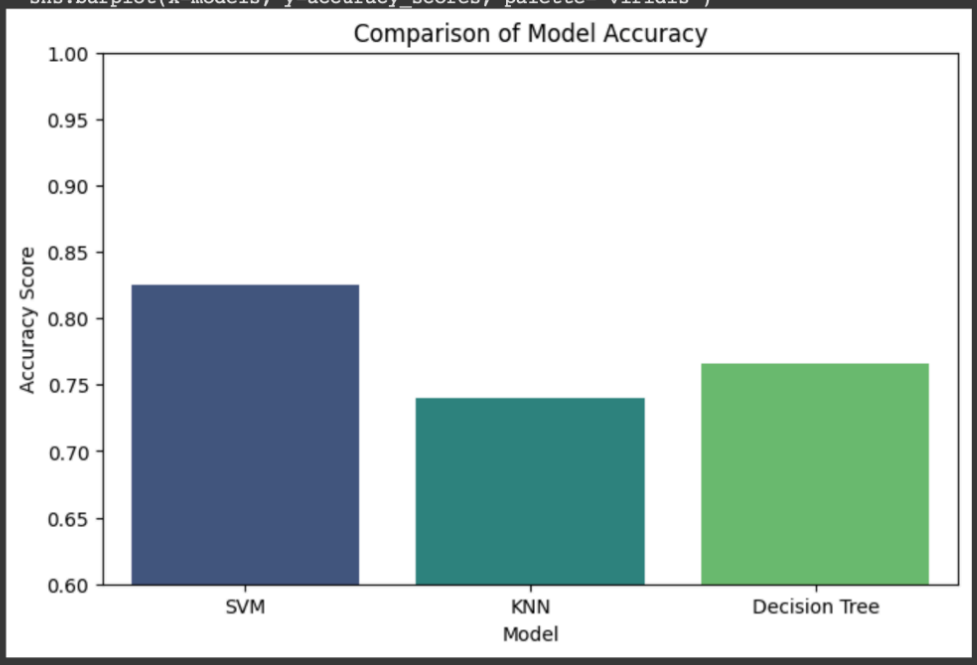
From the above heatmap we can say that:

- **Strong correlations** (above 0.7 or below -0.7) might indicate redundancy between features.
- As,**Glucose** and **Outcome (Diabetes)** are highly correlated, it suggests glucose is a strong predictor.
- Features with **low correlation (<0.3)** with Outcome may contribute less to prediction.
- The two independent variables (**BMI & Skin Thickness**) have high correlation, one might be redundant.

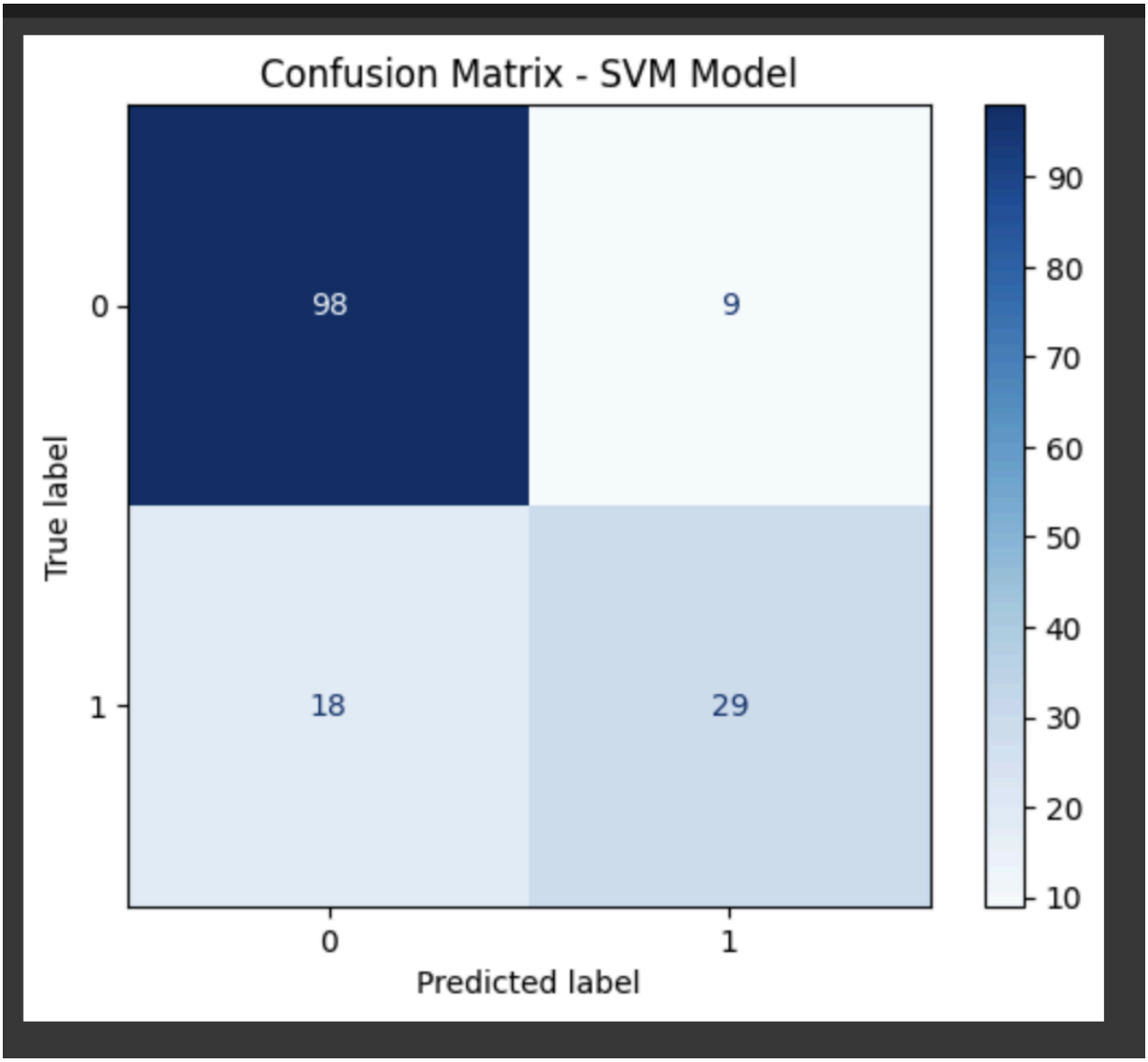
```
<ipython-input-23-a18b19c71ad1>:5: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x`

  sns.barplot(x=models, y=accuracy_scores, palette="viridis")
```



The above screenshot shows a bar graph that was used to visualize the accuracy of each three classification models i.e SVM, KNN, Decision Tree and lets us decide which among them gives the highest accuracy



The confusion matrix is used to check how many of the values are true positive ,true negative (i.e correctly predicted) and false positive and false negative (wrongly predicted) by using the SVM Model