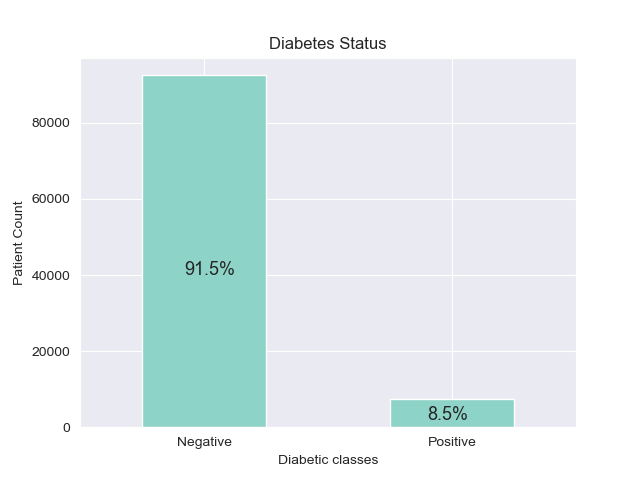
**A summary of the modelling process on a diabetes prediction dataset**

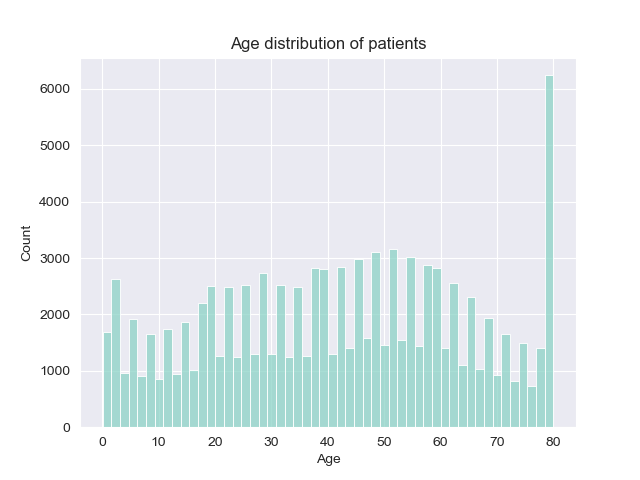
The dataset in question is a collection of medical and demographic data from 100,000 patients, with their diabetes status (positive or negative). The features represented in the dataset include:

**Diabetes Status:**

Fig 1

The dataset shows a heavy imbalance towards the negative class, with 91.5% of patients not diabetic and 8.5% of them diabetic.

**Age**:

 Fig 2

The dataset showed a fairly uniform distribution until the extreme end of the age feature, where approximately 6% of all recorded patients were age 80. Typically, diabetes is commonly diagnosed in older sections of the population and the dataset shows an over-representation of older patients(6% of patients alone are age 80). This occurrence is significant and should be taken into account.

**Gender**:

 Fig 3

The dataset showed that majority of the patients on record(almost 6 out of every 10), were female.

**Hypertension status**:

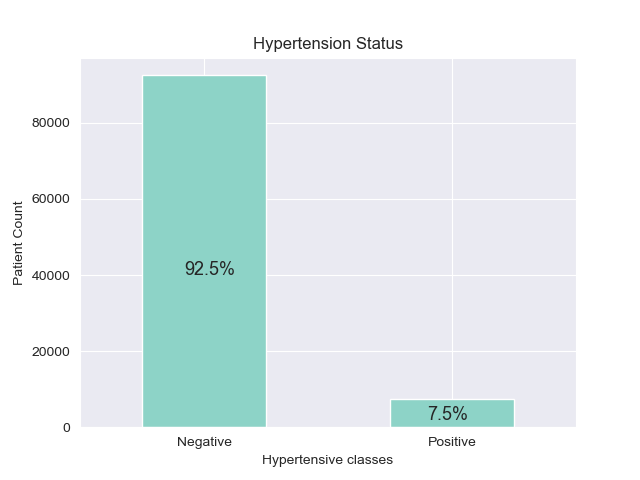


Figure 4

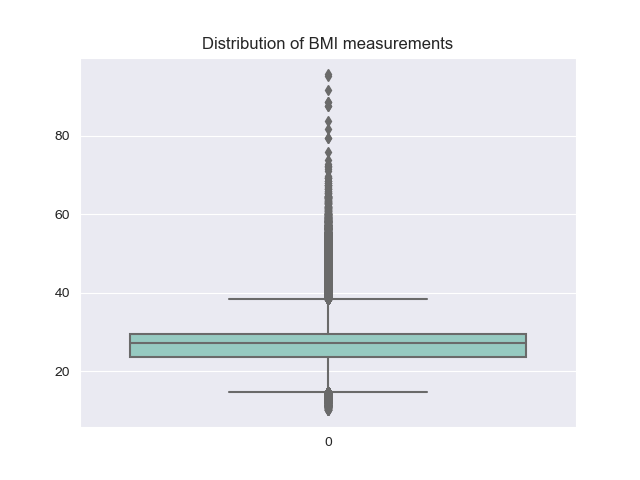
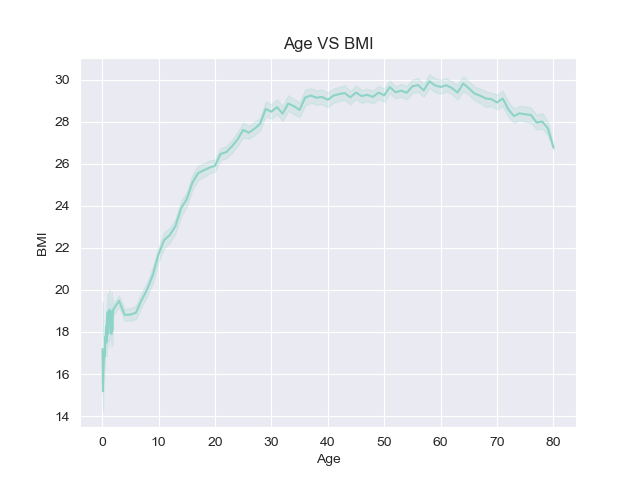
**BMI and Age VS BMI**:   
 

Figure 5 Figure 6

The plot on the left shows that the BMI feature is heavily skewed to the left and has the presence of a signficant number of outliers. The plot on the right shows a positive relationship between the age of patients and their BMI measurements. That is, the older the patient, the more likely their BMI measurement is higher. This makes intuitive sense as individuals in older age groups generally have more sedentary lifestyles, and by implication, may have increased BMI measurements.

**Haemoglobin A1c level**:

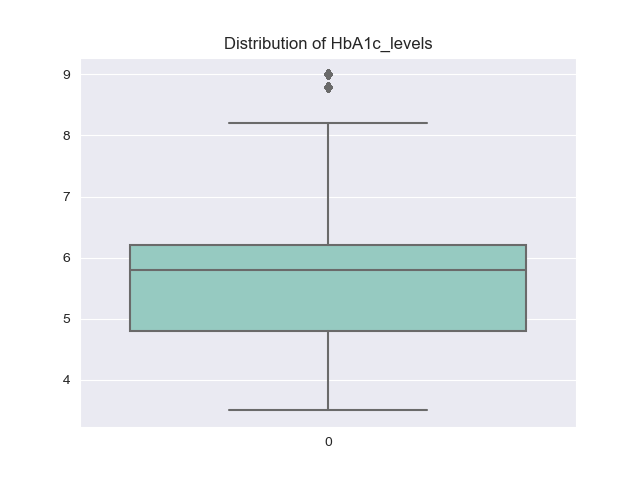


Figure 7

Higher Haemoglobin A1c levels show poor blood sugar control and an increased diabetic risk. Most of the patients have lower HbA1c levels, with very few occurences of outlier values.

Smoking Histories:

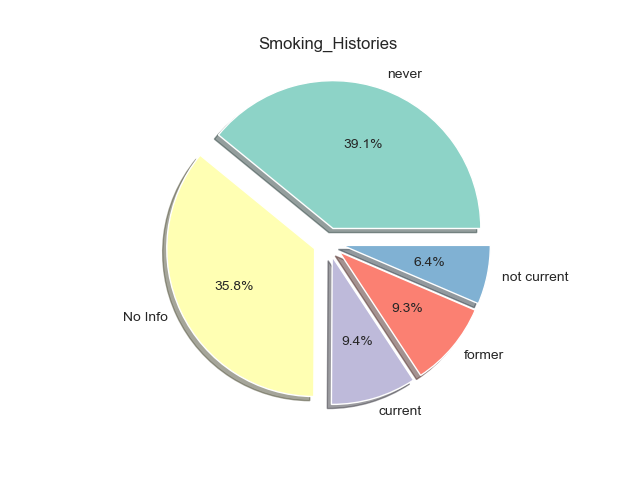


Figure 8

A significant portion of the patients on record have no experience smoking. However, close to a quarter of patients(24.1%) are either current smokers or have prior experience smoking. In general, smoking tends to complicate diabetes and is a risk factor.

**Correlation matrix**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Age | Hypertension | Heart Disease | BMI | HbA1c level | Blood glucose level | Diabetes  status |
| Age | 1.0 | 0.250519 | 0.234056 | 0.335949 | 0.101124 | 0.110115 | 0.257103 |
| Hypertension | 0.250519 | 1.0 | 0.121431 | 0.146010 | 0.082473 | 0.085471 | 0.199581 |
| Heart Disease | 0.234056 | 0.121431 | 1.0 | 0.055915 | 0.066585 | 0.069373 | 0.170365 |
| BMI | 0.335949 | 0.146010 | 0.055915 | 1.0 | 0.084224 | 0.094575 | 0.214777 |
| HbA1c level | 0.101124 | 0.082473 | 0.066585 | 0.084224 | 1.0 | 0.167900 | 0.401879 |
| Blood glucose level | 0.110115 | 0.085471 |  | 0.094575 | 0.167900 | 1.0 | 0.419316 |
| Diabetes  status | 0.257103 | 0.199581 | 0.170365 | 0.214777 | 0.401879 | 0.419316 | 1.0 |

**Basic statistical values of continuous features**

|  |  |  |
| --- | --- | --- |
| Feature | Mean | Standard Deviation |
| Age | 41 years | 22.5 years |
| HbA1c\_level | 5.5 | 1.07 |
| BMI | 27.3 | 6.63 |

**Model selection and results.**

The model which performed best on the data was a Random Forest Classifier model, which had the following results:

Accuracy Score (Training):0.9720625 Accuracy Score (Test):0.97135

Precision Score: 0.99 Recall Score: 0.66

F1 Score: 0.798

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1-Score |
| Negative class | 0.97 | 1.00 | 0.98 |
| Positive class | 1.00 | 0.67 | 0.80 |

**Confusion Matrix Results**

