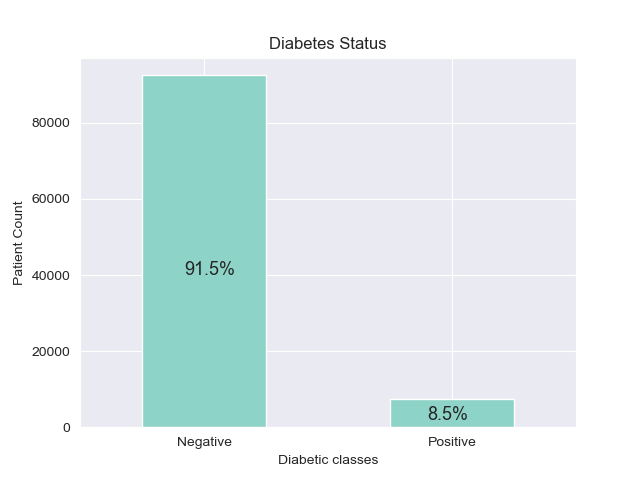
**A summary of the modelling process on a diabetes prediction dataset**

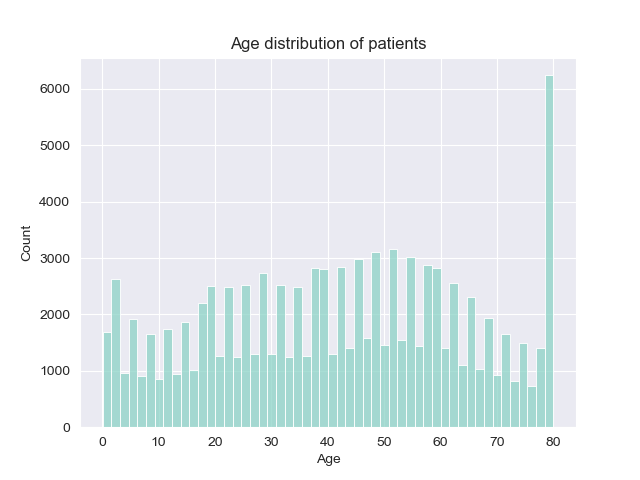
The dataset in question is a collection of medical and demographic data from 100,000 patients, with their diabetes status (positive or negative). The features represented in the dataset include:

**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**:

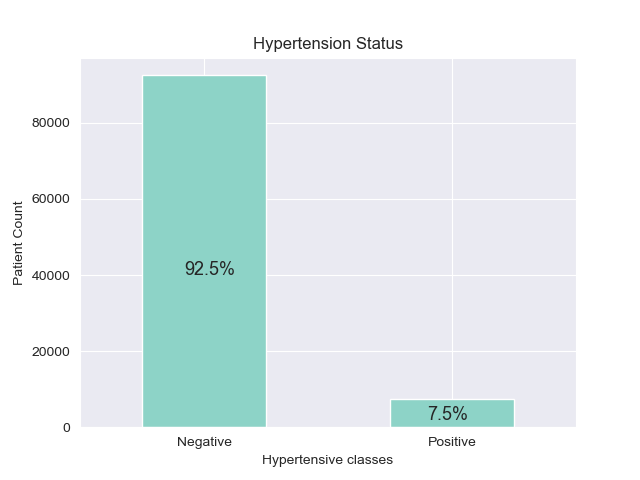


Figure 4

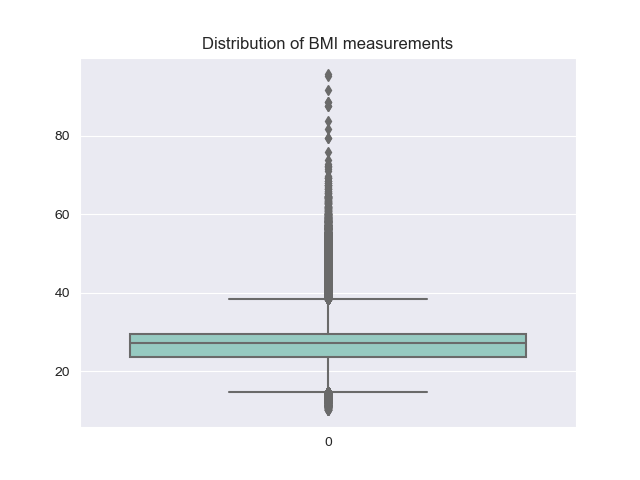
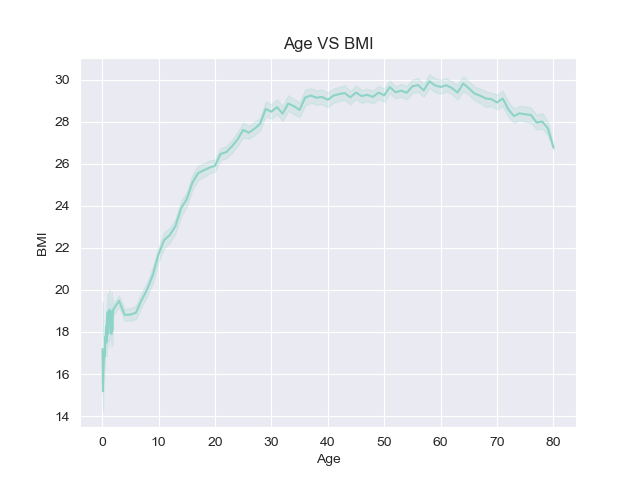
**BMI and Age VS BMI**:   
 

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**:

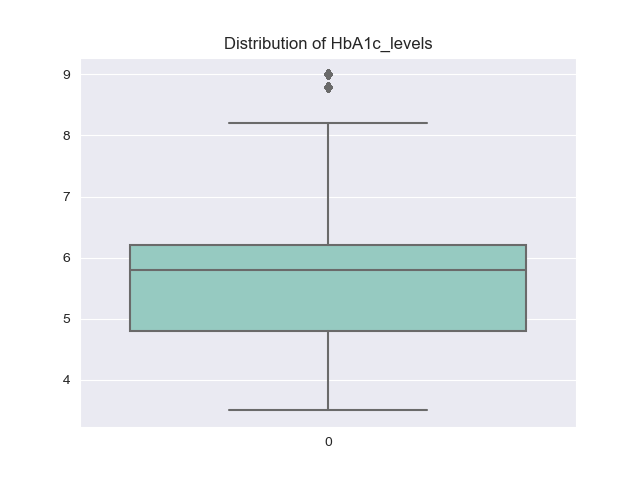


Figure 7

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**:

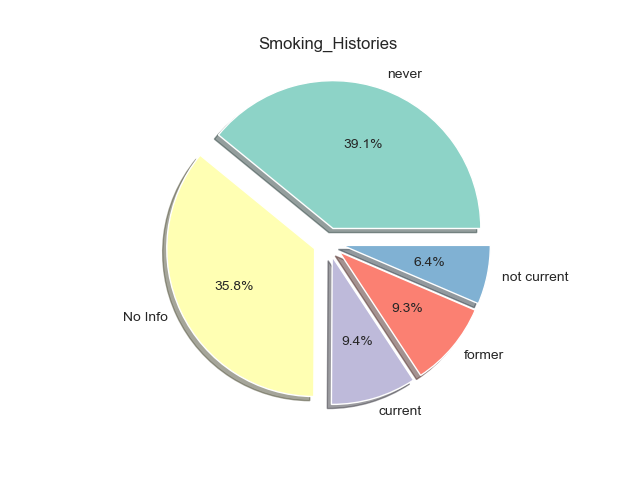


Figure 8

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 of the dataset**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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 |

**Insights into the model selection process**

1. **Addressing the imbalance of the target feature**: The dataset is heavily imbalanced, with 91.5% of patients in the negative class, and 8.5% in the positive class. To address this, Oversampling techniques were used(as against under sampling, to prevent information loss), to augment the data. Upon evaluation, chosen models performed worse on the Oversampled dataset when compared to the base training data. This also influenced the choice of an ensemble model, as ensemble models generally perform well on imbalanced datasets.
2. **Concerning Variations across groups**:

For gender differences, there were no significant differences in feature distributions owing to differences in sex. The tables below illustrate this finding:

|  |  |  |  |
| --- | --- | --- | --- |
| Female | Hypertension | Heart\_Disease | Diabetes |
| Negative | 92.8% | 97.3% | 92.3% |
| Positive | 7.16% | 2.67% | 7.7% |

Smoking History:

* Never 0.428798
* No Info 0.336453
* Current 0.086385
* Former 0.081534
* Not current 0.066829

|  |  |  |  |
| --- | --- | --- | --- |
| Male | Hypertension | Heart\_Disease | Diabetes |
| Negative | 92.06% | 94.25% | 90.2% |
| Positive | 7.93% | 5.74% | 9.7% |

Smoking History:

* No Info 0.388849
* Never 0.337630
* Former 0.110500
* Current 0.102052
* Not current 0.060970

The dataset did not suggest enough eveidence to show that varations in the features or indeed the target variable were sufficiently influenced by the sex of the patient. The proportional distributions of the features were largely similar across across the two genders.

1. **Dealing with Outliers**:

To deal with outliers, two approaches were considered:

* Log transformation of the relevant features
* Capping the features.

To deal with outliers, specifically in the BMI feature, a logarithmic transformation was applied and then the Standard Scaler was used. The reasons for this choice were as follows:

* The dataset is already heavily imbalanced in favour of the positive class. Any further cap set on the data may result in the loss of information on the negative class, which is already at a premium.
* Taking only the outliers in the BMI feature as a separate set, they contain nearly 25% of the negative class, which reinforces the point made above.

1. **Addressing the over-representation of 80 year old patients:**

To deal with the over-representation of 80 year olds in the dataset, a Stratified Shuffle Split, instead of a random split was employed in seperating the data into training and test sets. This way, the over-representation of this age group is accounted for. This split enables that the ages of the patients are adequately represented in both the training and the test sets.

**Model selection and results.**

The table below is a summary of implemented algorithms and their results:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy Score (Training) | Accuracy Score  (Test) | Precision  Score | Recall Score | F1 Score | ROC Score |
| Random Forest Classifier | 0.9720 | 0.97135 | 0.99 | 0.67 | 0.798 | 0.832 |
| Decision Tree Classifier | 0.9720 | 0.97125 | 0.99 | 0.66 | 0.797 | 0.8316 |
| Support Vector Classifier | 0.964 | 0.963 | 0.994 | 0.573 | 0.727 | 0.783 |
| Logistic Regression | 0.96025 | 0.9591 | 0.71 | 0.61 | 0.718 | 0.80 |

To generate the best model, Grid Search Cross Validation was employed to determine the best combination of hyper-parameters to get the best model performance.

Interestingly, the Random Forest and Decision Tree models have similar scores. The difference between the two models is marginal, however, the Random Forest Classifier is chosen due to the use case of the model in question, as marginal differences which may result in misdiagnosis, could be damaging.

**Feature Importances of the chosen model**

|  |  |
| --- | --- |
| Feature | Score |
| Age | 0.0150 |
| BMI | 0.0153 |
| HbA1c measurements | 0.581 |
| Blood glucose level | 0.379 |
| Smoking History | 0.0001 |
| Gender (Male) | 0.000 |
| Gender (Female) | 0.000 |
| Hypertension | 0.0012 |
| Heart Disease | 0.0069 |

**Confusion Matrix Results (Decision Tree Classifier)**

