# Insurance charges prediction

1) Dataset consists of 1338 records
2) 6 columns
3) Column names are
    a) Age
    b) Sex
    c) Bmi
    d) Children
    e) Smoker
    f) charges
4) Age and children fields are integer
5) Bmi and charges are float
6) Sex and smoker are objects
7) All the fields are non-null
8) There is no null values for all the columns
9) Data preprocessing:
    a) Convert the “sex” object column to “sex_male” 0 or 1. 0 means female and 1 means male. Is an int field.
    b) “Smoker” object column converted to “smoker_yes” 0 means non smoker and 1 means smoker. Is an int field.
    c) Created a “independent” data set with all the columns without “charges”
    d) Created a “dependent” data set with a “charges” column.

10) Different models are tried and their R-Score values are mentioned below.

11)Multilinear Regression:

```
from sklearn.linear_model import LinearRegression
regressor=LinearRegression()
regressor.fit(X_train,y_train)
```

```
LinearRegression
LinearRegression()
```

```
y_pred=regressor.predict(X_test)
```

```
from sklearn.metrics import r2_score
r_score=r2_score(y_test,y_pred)
```

```
r_score
```

```
0.7894790349867009
```

12) Support Vector machine(SVM):

Tried with various hyper parameters and results are below.

| Sno | Kernel | C | R-Score |
|---|---|---|---|
| 1 | linear | 10 | 0.00 |
| 2 | linear | 100 | 0.54 |
| 3 | linear | 500 | 0.63 |
| 4 | linear | 1000 | 0.63 |
| 5 | linear | 2000 | 0.69 |
| 6 | linear | 5000 | 0.76 |
| 7 | linear | 10000 | 0.74 |
| 8 | rbf | 10 | -0.08 |
| 9 | rbf | 100 | -0.12 |
| 10 | rbf | 500 | -0.12 |
| 11 | rbf | 1000 | -0.12 |

| | | | |
|---|---|---|---|
| 12 | rbf | 2000 | -0.11 |
| 13 | rbf | 5000 | -0.07 |
| 14 | rbf | 10000 | -0.02 |
| 15 | sigmoid | 10 | -0.09 |
| 16 | sigmoid | 100 | -0.12 |
| 17 | sigmoid | 500 | -0.46 |
| 18 | sigmoid | 1000 | -1.67 |
| 19 | sigmoid | 2000 | -5.62 |
| 20 | sigmoid | 5000 | -31.57 |
| 21 | sigmoid | 10000 | -119.52 |
| 22 | poly | 10 | -0.09 |
| 23 | poly | 100 | -0.10 |
| 24 | poly | 500 | -0.08 |
| 25 | poly | 1000 | -0.06 |
| 26 | poly | 2000 | 0.00 |
| 27 | poly | 5000 | 0.15 |
| 28 | poly | 10000 | 0.35 |

- Best SVM model Kernel has “Linear” and C value is 5000 and R-Score is 0.76
- But is less than the multi linear regression model and its R-Score has 0.78

13) Decision tree with various hyper parameters

| Sno | Criterion | Splitter/Estimators | Max_features | R-Score |
|---|---|---|---|---|
| 1 | squared_error | best | sqrt | 0.70 |
| 2 | squared_error | best | log2 | 0.69 |
| 3 | squared_error | best | None | 0.70 |
| 4 | squared_error | random | sqrt | 0.75 |
| 5 | squared_error | random | log2 | 0.61 |
| 6 | squared_error | random | None | 0.70 |
| 7 | friedman_mse | best | sqrt | 0.69 |
| 8 | friedman_mse | best | log2 | 0.66 |

| | | | | |
|---|---|---|---|---|
| 9 | friedman_mse | best | None | 0.68 |
| 10 | friedman_mse | random | sqrt | 0.62 |
| 11 | friedman_mse | random | log2 | 0.73 |
| 12 | friedman_mse | random | None | 0.73 |
| 13 | absolute_error | best | sqrt | 0.75 |
| 14 | absolute_error | best | log2 | 0.71 |
| 15 | absolute_error | best | None | 0.67 |
| 16 | absolute_error | random | sqrt | 0.70 |
| 17 | absolute_error | random | log2 | 0.66 |
| 18 | absolute_error | random | None | 0.71 |
| 19 | poisson | best | sqrt | 0.55 |
| 20 | poisson | best | log2 | 0.71 |
| 21 | poisson | best | None | 0.73 |
| 22 | poisson | random | sqrt | 0.55 |
| 23 | poisson | random | log2 | 0.69 |
| 24 | poisson | random | None | 0.64 |

- Best model is 0.75
  - 1) squared error , random and sqrt
  - 2) absolute error , best and sqrt
  - But both are less then the multi linear regression model and its R-Score has 0.78

Random Forest:

| Sno | Criterion | Splitter/Estimators | Max_features | R-Score |
|---|---|---|---|---|
| 1 | squared_error | 10 | sqrt | 0.8520006347 |
| 2 | squared_error | 10 | log2 | 0.8520006347 |
| 3 | squared_error | 10 | | 0.8330304134 |
| 4 | squared_error | 100 | sqrt | 0.8710271903 |
| 5 | squared_error | 100 | log2 | 0.8710271903 |
| 6 | squared_error | 100 | | 0.8538307913 |
| 7 | friedman_mse | 10 | sqrt | 0.8502777994 |
| 8 | friedman_mse | 10 | log2 | 0.8502777994 |

| | | | | |
|---|---|---|---|---|
| 9 | friedman_mse | 10 | | 0.8331662678 |
| 10 | friedman_mse | 100 | sqrt | 0.8710544016 |
| 11 | friedman_mse | 100 | log2 | 0.8710544016 |
| 12 | friedman_mse | 100 | | 0.8540518935 |
| 13 | absolute_error | 10 | sqrt | 0.8574290081 |
| 14 | absolute_error | 10 | log2 | 0.8574290081 |
| 15 | absolute_error | 10 | | 0.8350635553 |
| 16 | absolute_error | 100 | sqrt | 0.8710685856 |
| 17 | absolute_error | 100 | log2 | 0.8710685856 |
| 18 | absolute_error | 100 | | 0.8520093621 |
| 19 | poisson | 10 | sqrt | 0.8544955286 |
| 20 | poisson | 10 | log2 | 0.8544955286 |
| 21 | poisson | 10 | | 0.831399104 |
| 22 | poisson | 100 | sqrt | 0.8680156985 |
| 23 | poisson | 100 | log2 | 0.8680156985 |
| 24 | poisson | 100 | | 0.8526334259 |

- absolute_erro, 100 has estimators and sqrt/log2 has the **0.87 is the R-Score is highest**

**Conclusion:**

- **Best model is Random forest with** absolute_erro, 100 has estimators and sqrt/log2 has the **0.87 is the R-Score is highest**
- **Selected this model and deployment file created**
- **Predicted insurance charges from this model.**