भारतीय सूचना प्रौद्योगिकी संस्थान, नागपुर
Indian Institute of Information Technology, Nagpur
An Institution of National Importance By An Act of Parliament

Big Data Analysis Report

Under the guidance of: Dr Shishupal Kumar

Submitted by

Soumalya Roy - BT20CSE058

Mridul Gupta - BT20CSE092

1. Introduction:

Medical insurance charges are influenced by various factors such as age, gender, body mass index (BMI), number of children, smoking habits, and region of residence. In this project, we aim to predict the medical insurance charges using a linear regression model with polynomial expansion. We will use the Medical Cost Personal Datasets available on Kaggle to train and test our model.

2. Data Set Used for the Problem:

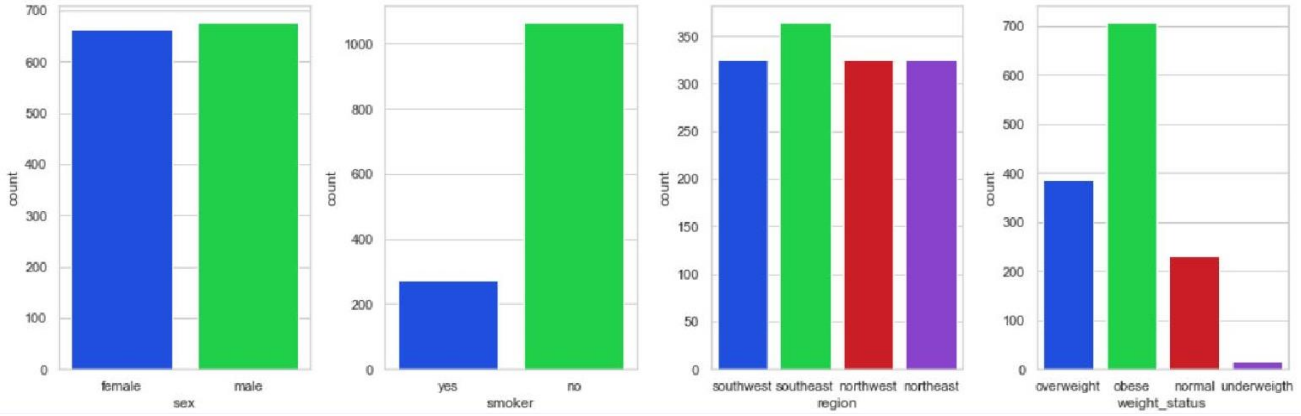
The dataset used in this project is the Medical Cost Personal Datasets obtained from Kaggle. The dataset consists of information about the medical and personal details of patients, as well as the charges billed by the hospital. The dataset contains 1338 observations with 7 features as described below:

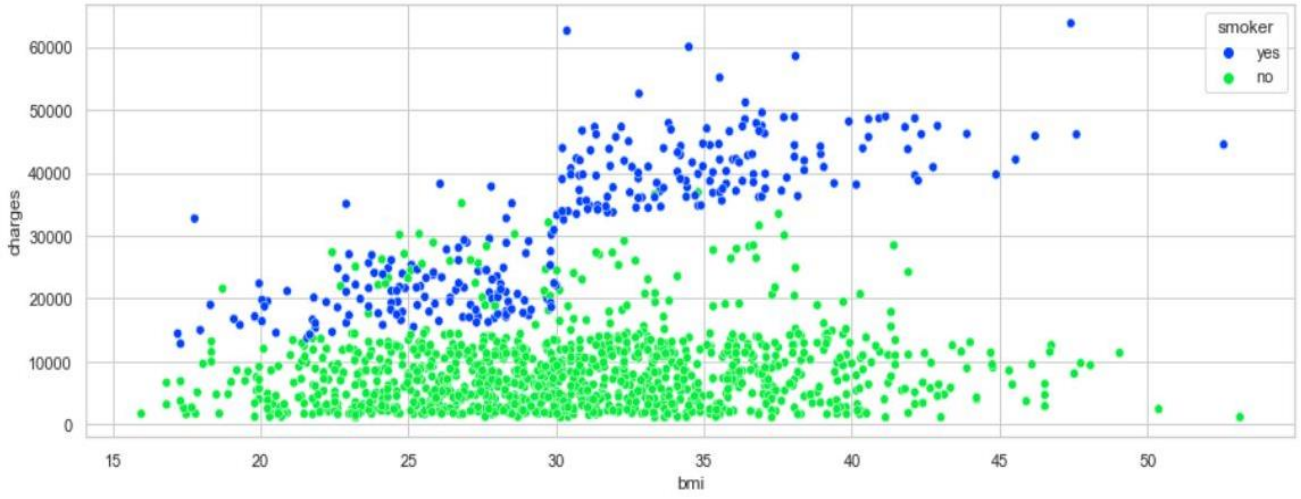
1. age: age of primary beneficiary
2. sex: gender of primary beneficiary (male = 0, female = 1)
3. bmi: body mass index of primary beneficiary
4. children: number of children covered by health insurance
5. smoker: smoking status of primary beneficiary (yes = 1, no = 0)
6. region: region where the beneficiary resides (northeast = 0, northwest = 1, southeast = 2, southwest = 3)
7. charges: individual medical costs billed by health insurance

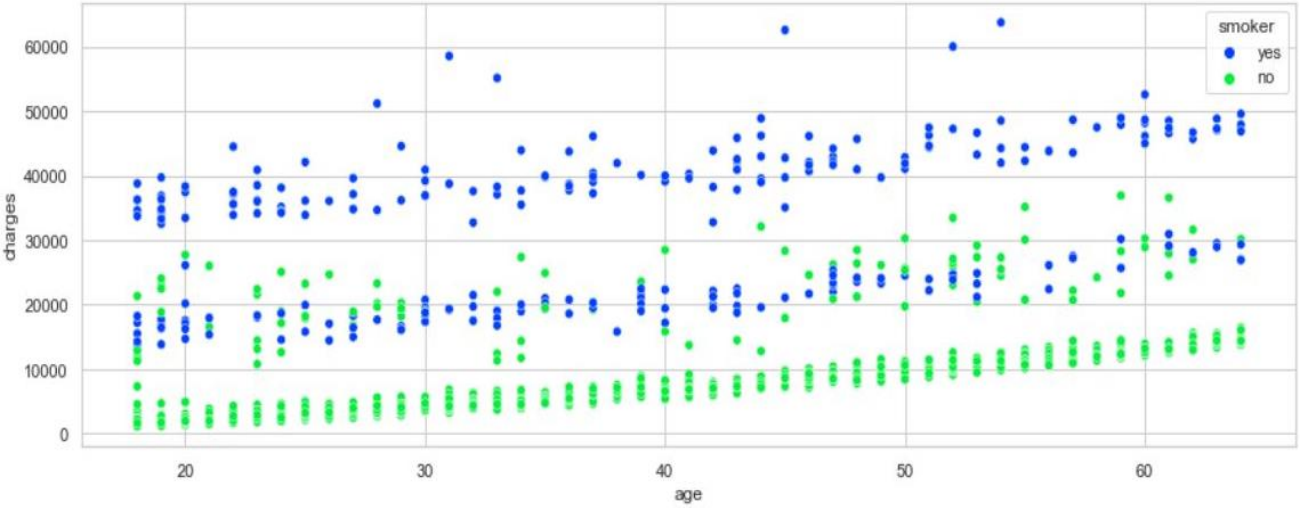
The dataset is a combination of categorical and numerical data. The 'sex', 'smoker', and 'region' features are categorical while 'age', 'bmi', 'children', and 'charges' features are numerical.

This dataset is relevant for predicting medical charges of patients based on their personal and medical information. This information can be useful for insurance companies to determine the premiums for their customers based on their health risks and medical history.

The dataset is relatively clean and free of missing values, making it suitable for machine learning analysis.


count
sex
female
male
smoker
yes
no
region
southwest
southeast
northwest
northeast
weight_status
overweight
obese
normal
underweigth


smoker
yes
no
charges
bmi


smoker
yes
no
charges
age

3. Method to solve the problem:

We will follow the following steps to solve the problem of predicting medical insurance charges:

- Data Preparation: We will import the dataset into a PySpark dataframe and perform data cleaning, exploration, and preprocessing. This will include handling missing values, encoding categorical variables, and scaling the numerical features.

- Feature Engineering:

a. Encoding sex, region, & smoker attribute using Vector Assembler in PySpark.

b. Conversion of age,bmi,children and charges to float values.

```
#Conversion of required string categories into Float Categories
int_cols = ["age", "bmi", "children","charges"]
for col_name in int_cols:
    df_pyspark = df_pyspark.withColumn(col_name,
col(col_name).cast("float"))
df_pyspark.show()
```

c. Assembling features together to gather for a target (charges) attribute.

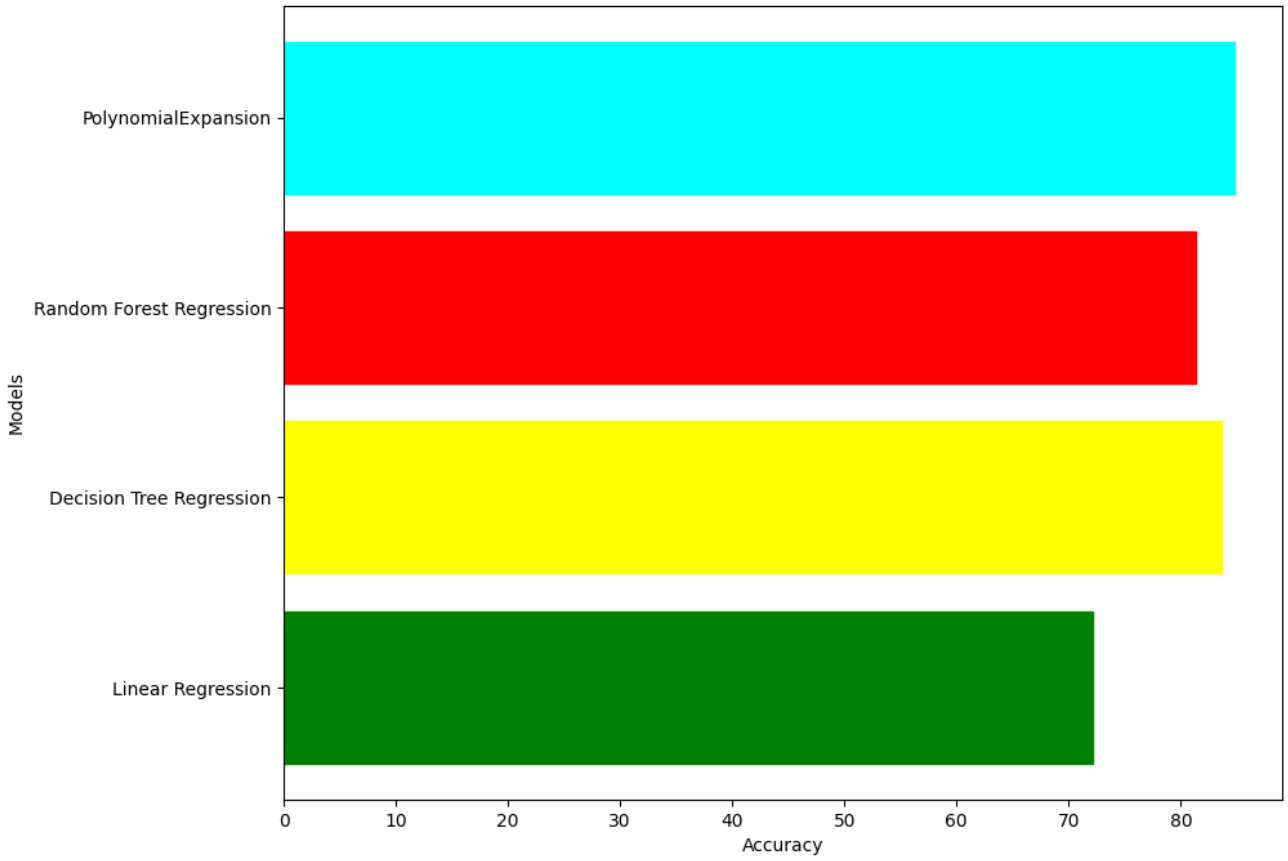
```
from pyspark.ml.feature import StringIndexer, VectorAssembler

indexer = StringIndexer(inputCols=["sex", "smoker", "region"],
outputCols=["sex_indexed", "smoker_indexed", "region_indexed"])

# Fit and transform the DataFrame
df_transformed = indexer.fit(df_pyspark).transform(df_pyspark)

# Define the input columns for the VectorAssembler
input_cols = ['age', 'sex_indexed', 'bmi', 'children', 'smoker_indexed',
'region_indexed']

# Create the VectorAssembler and transform the DataFrame
assembler = VectorAssembler(inputCols=input_cols, outputCol="features")
df_transformed = assembler.transform(df_transformed)

# Select the transformed column
df_transformed.select('features').show()
df_transformed.show()
```

We will use polynomial expansion to create higher-degree features from the original features. This will help us capture nonlinear relationships between the features and the target variable.

```
df = spark.createDataFrame([data])
df.describe

input_cols = ['age', 'sex_indexed', 'bmi', 'children', 'smoker_indexed',
'region_indexed']

# Create the VectorAssembler and transform the DataFrame
assembler = VectorAssembler(inputCols=input_cols, outputCol="features")
df_new_data = assembler.transform(df)


df_new_finalize = df_new_data.select("features")
#df_new_finalize.show()

poly_exp = PolynomialExpansion(degree=5, inputCol="features",
outputCol="poly_features")
df_newff = poly_exp.transform(df_new_finalize)

prediction = plr_model.transform(df_newff)

# Extract the predicted charges value
charges = prediction.select("prediction").collect()[0][0]

print("Predicted charges: $%.2f" % charges)
```

- Model Training: We will split the dataset into training and testing sets, and train a linear regression model on the training set using the expanded features. We will use PySpark's MLlib library for this purpose. A short description on all the four different Algorithm's we applied.

- Model Evaluation: We will evaluate the performance of the trained model on the testing set using evaluation metrics such as Mean Squared Error , Root mean squared error (RMSE) and R-squared.



- Prediction: We will use the trained model to predict the medical insurance charges for new data points and Incur charges of premium for any new Customer that wants to Purchase the Medical Insurance.

4. Conclusion:

In this project, we used a linear regression model with polynomial expansion to predict medical insurance charges. We used PySpark's MLlib library to perform the data preprocessing, feature engineering, and model training. We achieved a reasonable RMSE value of 4540.874 and R-squared value of 0.848 on the testing set, which indicates that the model is able to capture the underlying relationships between the features and the target(charges) variable. This model can be used by insurance companies to estimate the charges for their customers based on their personal and medical information.

5. Future Work:

Some future work that can be done to improve the performance of the model are:

- Collect more data on other factors that can influence medical insurance charges such as lifestyle habits, family history, and pre-existing medical conditions.

- Based on the above collected data we are assuming that insurance premium to be paid is of the same amount but in reality different insurance premiums are paid for different insured amounts.

- Moreover with past data , eating habbits, pre-existing medical conditions we can determine whether the person is likely to fall for any disease quickly and the amount insured to be is required high thus determining from food habits about the pricing based on certain such factors.

6. References:

- The Medical Cost Personal Datasets available on Kaggle: https://www.kaggle.com/mirichoi0218/insurance

- PySpark's MLlib library documentation: https://spark.apache.org/docs/latest/ml-guide.html

- PySpark Tutorial on Youtube : https://www.youtube.com/watch?v=_C8kWso4ne4&t=2s