Objective:

The main purpose of this document is to identify the best model to predict the insurance premium for the individual based on the given criteria in the dataset.

Requirements & Decisions:

Project Name	Insurance Premium Predictor
Dataset Filename	insurance_pre.csv
Goal	To predict the insurance premium of an individual.

Problem Identification	
Stage1 (Domain Selection)	Machine Learning
Stage2 (Learning Method)	Supervised Learning
Stage3 (Data Type)	Regression

Dataset Info.	
No. of column	6
No. of rows	1338
INPUT Column Name	Age, Sex, BMI, Children, Smoker
OUTPUT Column Name	Charges

Research Values:

Based on request, the dataset was imported and the models are created using different algorithms in machine learning. The accuracy of the models are captured and documented below.

Pre-processing	
Data Type	Nominal
Method	One Hot Encoding
Purpose	Converting the string to numerical data

Algorithms used to predict	1.Mutiple Linear regression 2.Support Vector Machine 3.Decision Tree Regression 4.Random Forest Regression
Evaluation Metrics	r2_score

Algorithms and R2 Values:

The R2 value for **Multiple Linear Regression** is **0.789479**.

Multiple Linear regression	R2 Value	0.789479

The Best R2 value for **Support Vector Machine** is **0.877995** using the Hyper Factor parameter **C=10000.**

		kernel				**Hyper Factor (C)**
Support Vector Machine	R2 Value	linear	poly	rbf	sigmoid	
		-0.010102	0.038716	-0.083382	-0.075429	1
		0.462468	0.617956	-0.032273	0.039307	10
		0.628879	0.617956	0.320031	0.52761	100
		0.76493	0.856648	0.810206	0.28747	1000
		0.741423	0.859171	**0.877995**	-34.151535	10000

The Best R2 Value for **Decision Tree Regression** is **0.742155** using the Hyper Factor parameters **criterion='sqared_error', splitter='best', max_features='log2'.**

		splitter	**max_features**	***criterion***			
Decision Tree Regression	R2 Value			*squared_error*	*friedman_mse*	*absolute_error*	*poisson*
		best	*sqrt*	0.735031	0.69424	0.685038	0.738281
			log2	**0.742155**	0.650137	0.719249	0.693742
			None	0.69313	0.709051	0.670448	0.712466
		random	*sqrt*	0.67981	0.547566	0.699331	0.70492
			log2	0.73104	0.690269	0.617093	0.695381
			None	0.697368	0.700833	0.703366	0.68352

The Best R2 Value for **Random Forest Regression** is **0.873995** using the Hyper Factor parameters **criterion='absolure_error', bootstrap=TRUE, max_features='log2', n_estimators=100.**

Random Forest Regression	R2 Value	**n_estimators=10**	**bootstrap**	**max_features**	***criterion***			
					squared_error	*friedman_mse*	*absolute_error*	*poisson*
			TRUE	*sqrt*	0.854116	0.859664	0.854797	0.850702
				log2	0.856894	0.845378	0.849596	0.856124
				None	0.861066	0.833268	0.836587	0.827359
			FALSE	*sqrt*	0.839213	0.839919	0.825159	0.833108
				log2	0.838229	0.836644	0.827215	0.823331
				None	0.697185	0.695152	0.693856	0.73614
		n_estimators=100	**bootstrap**	**max_features**	***criterion***			
					squared_error	*friedman_mse*	*absolute_error*	*poisson*
			TRUE	*sqrt*	0.871757	0.871509	0.87085	0.869206
				log2	0.869616	0.872181	**0.873995**	0.868791
				None	0.856132	0.853654	0.853985	0.8501195
			FALSE	*sqrt*	0.846779	0.845727	0.844496	0.845195
				log2	0.845229	0.849785	0.841603	0.849873
				None	0.69947	0.70225	0.692502	0.733205
		n_estimators=1000	**bootstrap**	**max_features**	***criterion***			
					squared_error	*friedman_mse*	*absolute_error*	*poisson*
			TRUE	*sqrt*	0.872131	0.872363	0.873904	0.871963
				log2	0.872409	0.873226	0.873667	0.872535
				None	0.85634	0.85456	0.854506	0.855085
			FALSE	*sqrt*	0.847509	0.847112	0.846014	0.847939
				log2	0.848441	0.847606	0.84408	0.848821
				None	0.703602	0.702176	0.693536	0.732958

Research Observation:

The below table shows the best R2 value from each algorithm with the respective Hyper Factor parameters.

Algorithm	**Best R2 Value**	**Hyper Factor Parameter**
Multiple Linear regression	0.789479	-
Support Vector Machine	**0.877995**	**C=10000**
Decision Tree Regressor	0.742155	criterion='sqared_error', splitter='best', max_features='log2'
Random Forest Regressor	0.873995	criterion='absolure_error', bootstrap=TRUE, max_features='log2', n_estimators=100

Conclusion:

Based on the above research table the model created using the algorithm **Support Vector Machine** using the Hyper Factor parameter **C=10000** is having the higher R2 VALUE of **0.877995** and should be used for this project.