TECH LEAD 2025.1
Projeto Final

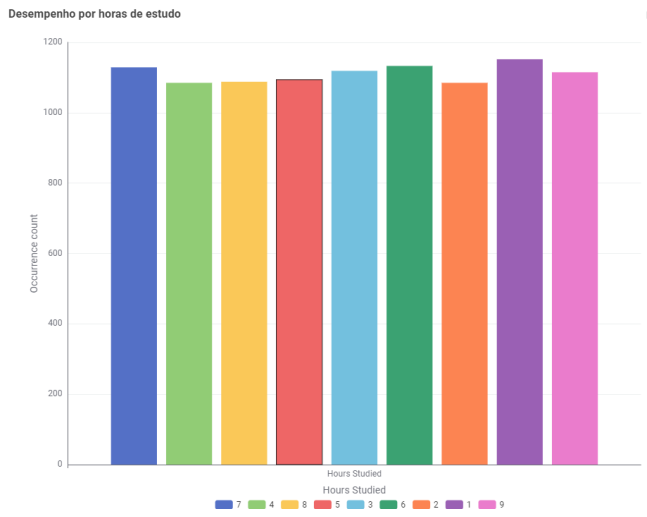
Disciplina: Data Science e IA
Aluno1: Sttiwe Washington F Sousa
E-mail1: swfs@cesar.school
Aluno2: Pedro William Bernardino
E-mail2: pwbcf@cesar.school
Student Performance: https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression

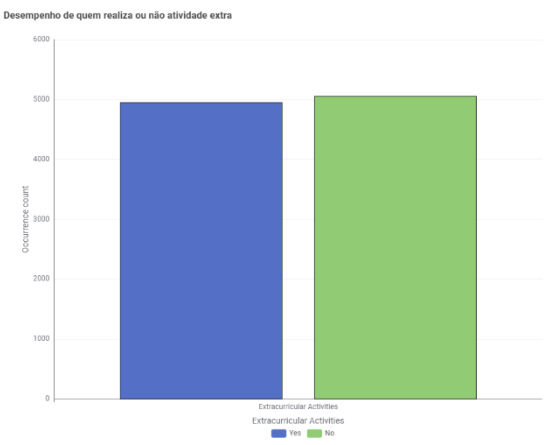
Inicialmente após ler o arquivo .csv com o CSV Reader podemos obter algumas informações sobre os dados. O arquivo contém um conjunto de dados de com 10 mil entradas, com as seguintes variáveis:

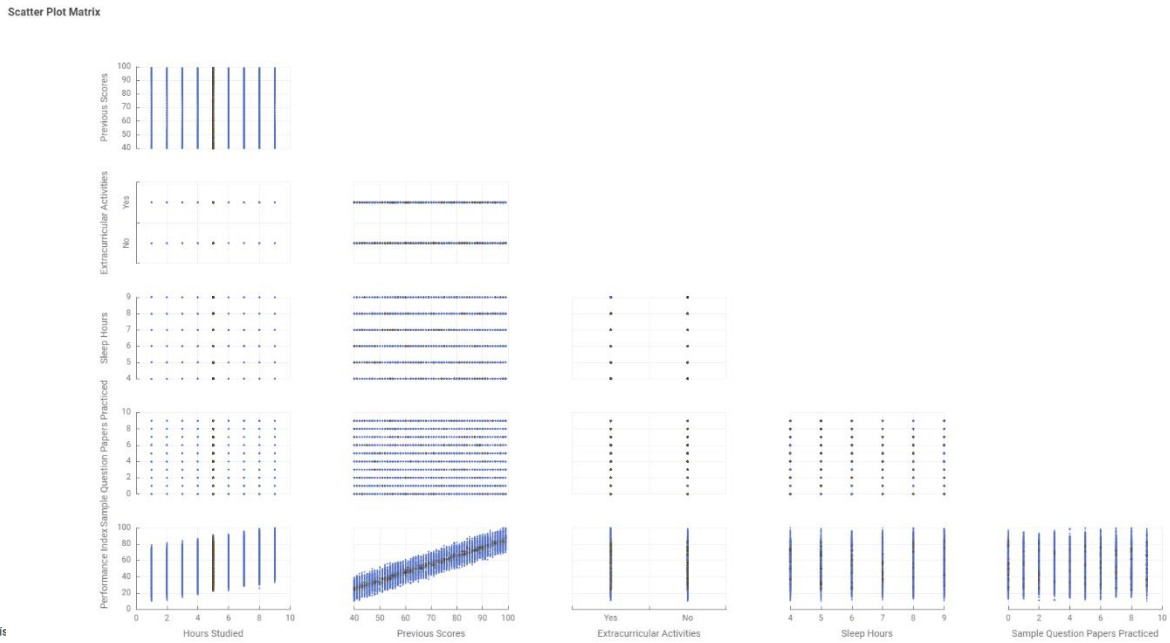
- ***Hours Studied***: O número total de horas dedicadas aos estudos por cada aluno.
- ***Previous Score***: As notas obtidas pelos alunos em provas anteriores.
- ***Extracurricular Activities***: Se o aluno participa de atividades extracurriculares (Sim ou Não).
- ***Sleep Hours***: O número médio de horas de sono que o aluno teve por dia.
- **Sample Question Papers Praticed**: O número de provas de exemplo que o aluno praticou.
- **Performance Index**: Média de desempenho de cada aluno.

Essa última variável será o nosso valor de referência para análise.

Iniciamos com algumas visualizações. Incluí uma View Bar Chat para analisar e compara e visualizar graficamente a relação do desempenho com as outras variáveis.







Estatís

Rows: 6

Name	# Unique values	Minimum	Maximum	25% Quantile	50% Quantile (Median)	75% Quantile	Mean	Mean Absolute Deviati...	Standard Deviation	Sum	10 most common values
Extracurricular Activities	2										No (5052; 50.52%), Yes (4948; 49.48%)
Performance Index	91	10	100	40	55	71	55.225	16.204	19.213	552,248	67 (187; 1.87%), 45 (185; 1.85%), 49 (18
Previous Scores	60	40	99	54	69	85	69.446	15.068	17.343	694,457	54 (216; 2.16%), 87 (208; 2.08%), 56 (19
Sleep Hours	6	4	9	5	7	8	6.531	1.489	1.696	65,306	8 (1804; 18.04%), 7 (1676; 16.76%), 6 (1
Sample Question Papers	10	0	9	2	5	7	4.583	2.492	2.867	45,833	6 (1059; 10.59%), 9 (1051; 10.51%), 3 (1
Hours Studied	9	1	9	3	5	7	4.993	2.231	2.589	49,929	1 (1152; 11.52%), 6 (1133; 11.33%), 7 (1

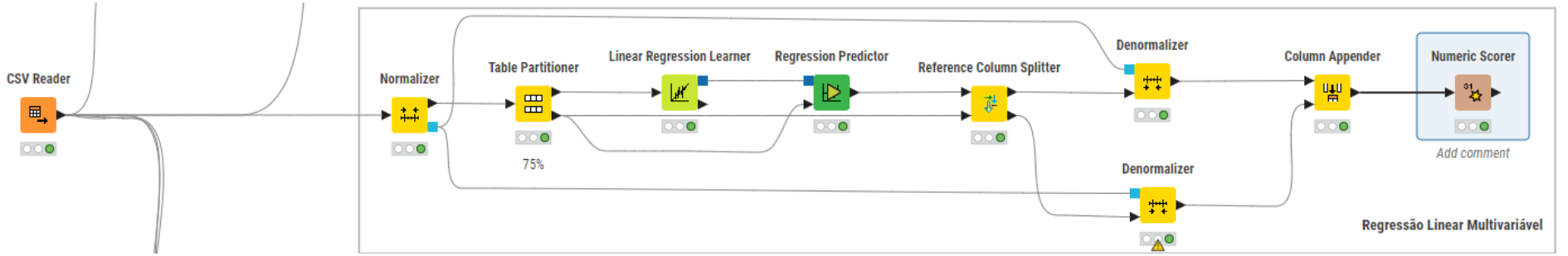
Posteriormente incluímos uma Linear Correlation para visualizar a correlação das variáveis com o valor de performance:

First column name *String*	Second column name *String*	Correlation value *Number (Float)*
Previous Scores	Performance Index	0.915
Hours Studied	Performance Index	0.374
Sleep Hours	Performance Index	0.048
Sample Question Papers Practiced	Performance Index	0.043

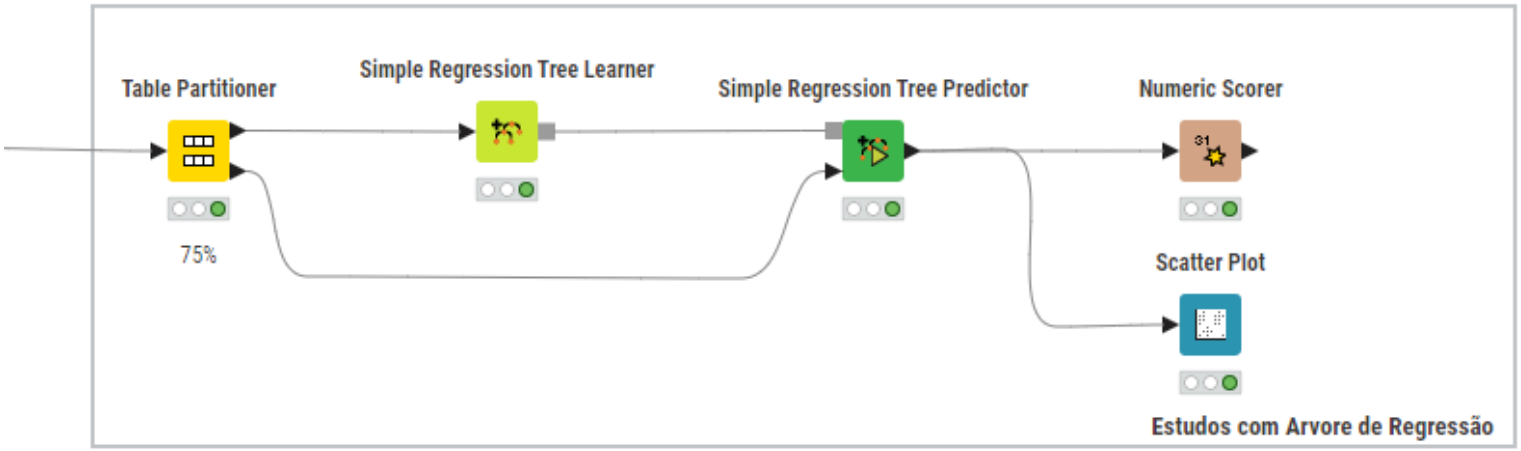
Com a primeira análise podemos observar que os alunos com bom desempenho possuem uma quantidade razoável de horas de sono. O histórico de notas do aluno que possui um bom desempenho, possui uma correlação muito alta no desempenho atual, apresentando assim uma regularidade na sua média.

Tentamos realizar a análise dos dados utilizando os exemplos visto em aula utilizando o a Regressão Linear Multivariável. Normalizamos os dados e incluímos o particionamento da tabela configurada em 75%, assim, utilizamos o Linear Regression Learner e o Regression Predictor, denomalizei os dados e incluí o Numeric Scorer. Não tivemos uma boa compreensão do por que o resultado ficou negativo em R^2 = -8.023.

Assim, segue imagem da utilização da regressão linear multivariável.



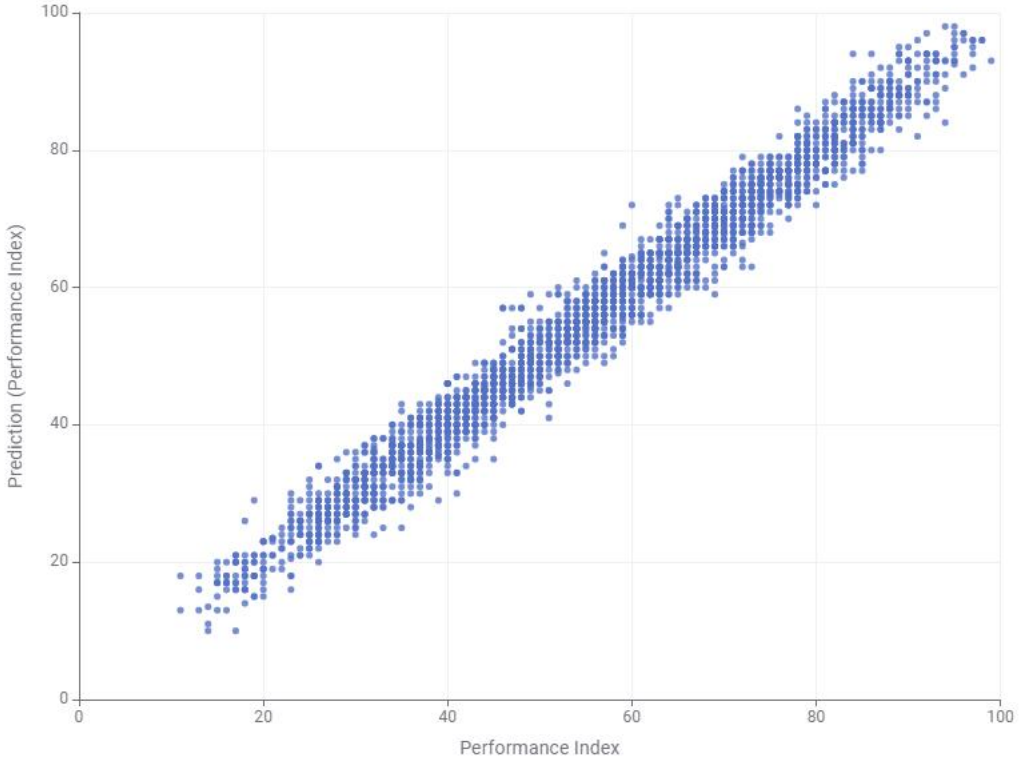
Realizamos algumas pesquisas no fórum kaggle e umas das soluções recomenda a utilização da Simple Regression Tree Learner. A solução utilizando a arvore de decisão para o problema de regressão linear nos trouxe melhores resultados. Assim, ficou a estrutura no Knime:



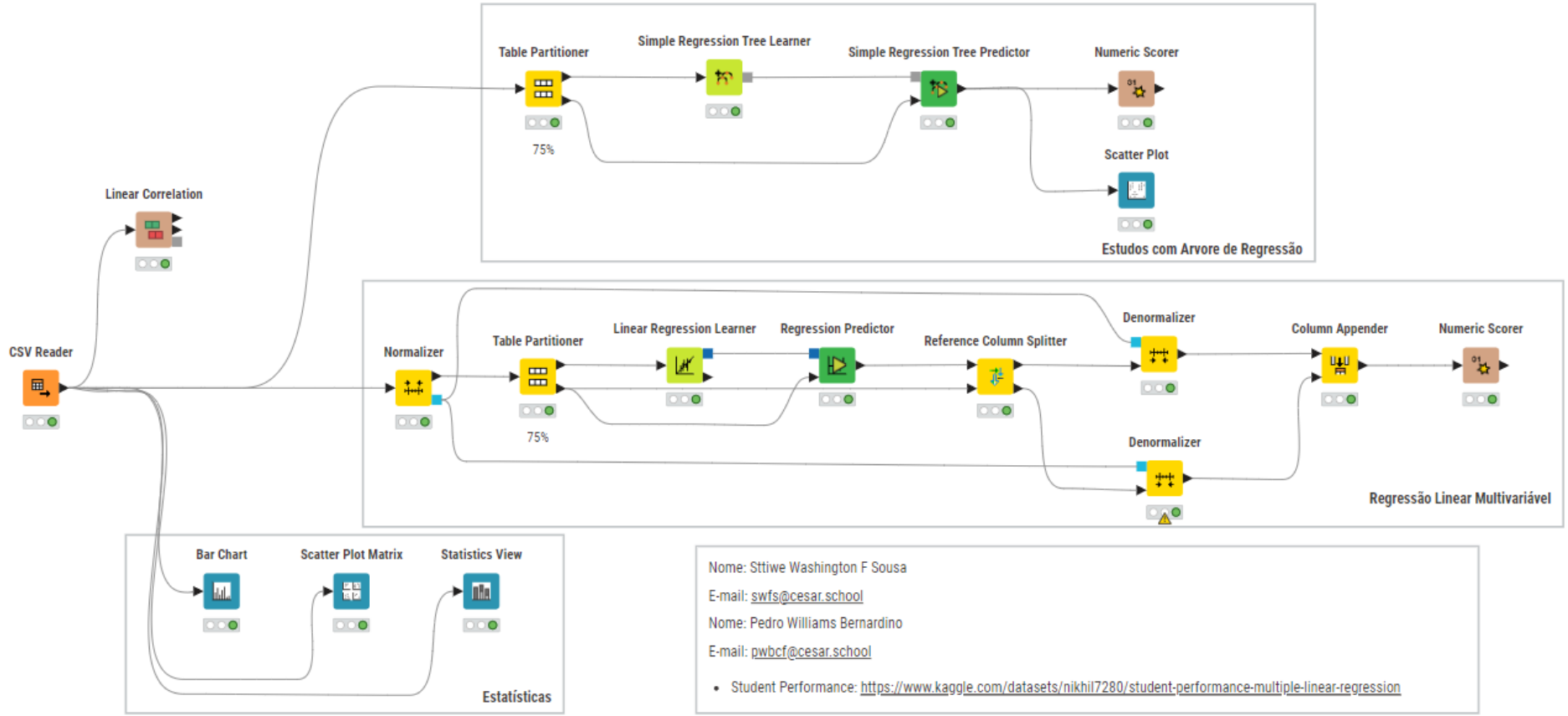
Obtivemos um resultado positivo tendo um índice de 97,5% da variação de saída.

- **R^2** = 0.975
- **Mean ansolute error** = 2.37

RowID	Prediction (Performance Index) *Number (Float)*
R^2	0.975
mean absolute error	2.37
mean squared error	9.264
root mean squared error	3.044
mean signed difference	-0.103
mean absolute percentage error	0.05
adjusted R^2	0.975



Assim ficou toda solução:



Nome: Sttiwe Washington F Sousa

E-mail: swfs@cesar.school

Nome: Pedro Williams Bernardino

E-mail: pwbcf@cesar.school

- Student Performance: https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression