Relatório
Projeto Final de Programação INF2102
2022.2

Nome do(a) Aluno(a)	André Luiz Farias Novaes
Número de Matrícula	2112511
Nome do(a) Orientador(a)	Hélio Côrtes Vieira Lopes
Trilha de *Open Science*	☐ Sim **X Não**
<u>Complexidade</u>: O programa atende aos requisitos de complexidade exigidos, em especial aos itens *tamanho* e *natureza*?	**X Sim** ☐ Não

Informações gerais

Natureza do Trabalho

O resultado do Projeto Final de Programação (PFP), nomeado *projeto*, é um programa devidamente especificado, projetado, desenvolvido e testado.

O projeto é uma extensão do algoritmo apresentado e desenvolvido em [1], [2], [3] e [4]. A *Programação Genética Econométrica* (ou EGP, sigla em inglês), um gerador de modelos parcimoniosos híbridos de Inteligência Artificial (IA) e econometria para previsão em dados puramente seccionais, foi implementada como extensão do pacote GPTIPS (https://sites.google.com/site/gptips4matlab/). A EGP gera soluções de acurácia elevada e competitivas frente aos benchmarks vigentes, com interpretabilidade de parâmetros. Neste projeto, a EGP foi estendida à previsão em dados de séries temporais (ST), em um esforço adicional de implementação e idealização do projeto, visto que STs envolvem, por definição, correlação serial temporal.

O programa atende aos requisitos de **utilidade** exigidos pois realiza um serviço de interesse potencial à comunidade científica de econometria, IA ou ambas, trazendo modelos parcimoniosos híbridos de IA e econometria para previsão em dados de séries temporais e dados puramente seccionais. A relevância da previsão de séries temporais à comunidade científica, industrial, comercial e, em sentido amplo, à sociedade, pode ser avaliada em [5].

O programa atende aos requisitos de **complexidade** exigidos pois atende aos sub-requisitos de **tamanho** e **natureza** do problema resolvido pelo programa.

O programa atende aos requisitos de **especificação** exigidos, ao responder às seguintes perguntas:

a. Qual é a **Finalidade** do programa? Gerar modelos parcimoniosos híbridos de IA e econometria para previsão, em dados de séries temporais e dados puramente seccionais, que sejam competitivos frente aos algoritmos *State-of-the-Art* (SOTA) e superiores aos algoritmos *benchmark*.

b. Quais **Características** deve possuir? A **Especificação de Requisitos do Software** responderá integralmente esta pergunta.

c. Quais **Funcionalidades** deve oferecer aos usuários? A **Especificação de Requisitos do Software** responderá integralmente esta pergunta.

1. Especificação

1.1. Escopo

1.2. Listagem de requisitos funcionais (ou *features*)

Serão descritos da seguinte forma: "[RFn] O software deve permitir que o [responsável pela ação] + [ação]", ou expressão similar.

[RF1] O software deve permitir que o usuário realize experimentos no âmbito de algoritmos híbridos em IA e Econometria.

[RF2] O software deve permitir que o usuário programe seus próprios experimentos.

[RF3] O software deve permitir que o usuário recrie os experimentos realizados especificamente em[1], [2], [3], [4] e [5].

1.3. Listagem de requisitos não funcionais

Serão descritos da seguinte forma: "[RNFn] O software deve permitir que o [responsável pela ação] + [ação] + [métrica]", ou expressão similar.

Requisitos da plataforma:

[RNF1] O software deve permitir que o usuário realize experimentos no âmbito de algoritmos híbridos em IA e Econometria dentro da plataforma MATLAB, já que o programa foi criado tendo como base o programa GPTIPS, desenvolvido em MATLAB.

[RNF2] O software deve permitir que o usuário programe seus próprios experimentos dentro da plataforma MATLAB.

[RNF3] O software deve permitir que o usuário recrie os experimentos realizados especificamente em[1], [2], [3], [4] e [5] dentro da plataforma MATLAB.
Requisitos do produto:

Requisitos do produto:

- de Usabilidade:

[RNF4] O software deve permitir que os usuários iniciantes em IA, Econometria ou ambos utilizem até 15% das funções do sistema após leitura da sessão 'Usuários Iniciantes' da Documentação ao Usuário.

[RNF5] O software deve permitir que os usuários intermediários em IA, Econometria ou ambos utilizem até 40% das funções do sistema após leitura da sessão 'Usuários Intermediários' da Documentação ao Usuário.

[RNF6] O software deve permitir que os usuários avançados em IA, Econometria ou ambos utilizem pelo menos 40% das funções do sistema após leitura da sessão 'Usuários Avançados' da Documentação ao Usuário.

- de Confiabilidade:
[RNF7] O software deve estar disponível aos usuários 99,99% das vezes.

- de Segurança:
[RNF8] O acesso aos dados deve ser protegido por senha alfanumérica de pelo 8 dígitos.

- de Desempenho:
[RNF9] O software deve processar a avaliação de um indivíduo à taxa de 1 avaliação/segundo.

- de Capacidade:
[RNF10] O software deve suportar pelo menos 4 usuários concorrentemente (impossibilidade de mais usuários concorrentes por limitações da MathWorks).

- de Portabilidade:
[RNF11] O software deve rodar nas plataformas Windows, Linux e Mac, exclusivamente.

Requisitos Organizacionais:

- de Entrega:
[RNF12] O software deve entregar um relatório de progresso dos experimentos a cada execução completa dos algoritmos implementados.

- de Implementação:
[RNF13] O software deve ser implementado na linguagem MATLAB.

Requisitos Externos:

- de Compatibilidade:
[RNF14] O software deve interagir com o GPTIPS.

- Éticas:
Não há.

- Legais:
[RNF15] O software deve ser distribuído e utilizado de acordo com as mesmas leis aplicadas ao MATLAB, distribuído pela MathWorks.

Requisitos de domínio:

[RD1] O software deve permitir aos usuários aqui.

[RD1] O treinamento dos modelos deve ocorrer antes da validação dos modelos.

[RD2] A validação dos modelos deve ocorrer antes do teste dos modelos.

[RD3] O usuário deve escolher o parâmetro k da validação cruzada do tipo *k-fold*.

[RD4] O usuário deve escolher o tipo de tarefa a resolver com o software: regressão ou classificação.

[RD5] O usuário deve escolher o tipo de dataset a aplicar os métodos: transversais/seccionais ou séries de tempo.

[RD6] O usuário deve ser capaz de alterar os percentuais dos dados atribuídos ao treino, validação e teste.

[RD7] O percentual de dados atribuído ao conjunto de validação não pode ser maior do que o percentual de dados atribuído ao conjunto de treino.

[RD8] O percentual de dados atribuído ao conjunto de teste não pode ser maior do que o percentual de dados atribuído ao conjunto de validação.

[RD9] O percentual de dados atribuído ao conjunto de validação deve ser de, no mínimo, 10%.

[RD10] O percentual de dados atribuído ao conjunto de teste deve ser de, no mínimo, 10%.

[RD11] As métricas de acurácia possíveis, para os experimentos envolvendo a tarefa de regressão, são: o *Root Mean Squared Error* (RMSE) e o R^2 Ajustado.

[RD12] A métrica de acurácia para os experimentos envolvendo a tarefa de classificação é: o percentual de classificações corretas, para datasets balanceados; e a Precisão e o Recall, para datasets desbalanceados..

1.4. Detalhamento (especificação) dos requisitos funcionais (ou das *features*)

1.4.1. Descrição textual simples

Não foi realizado.

1.4.2. Histórias de usuário (idealmente acompanhados de protótipos de telas e cenários de aceitação BDD)

Histórias de usuário serão descritas da seguinte forma: “[HUn] Como [usuário] eu quero [algo] para [finalidade]”, ou expressão similar

[HU1] Como [usuário] eu quero [treinar modelos híbridos de IA + Econometria] para [tarefas de regressão ou classificação em datasets de dados seccionais ou de séries de tempo].

Cenário BDD associado ao [HU1]:

Funcionalidade		
Como	usuário	
Quero	treinar modelos híbridos de IA + Econometria	
Para	tarefas de regressão ou classificação em datasets de dados seccionais ou de séries de tempo	
Cenário		
	Dado que	Estou no script/tela de seleção de tarefas e datasets.
	Quando	Seleciono uma tarefa (regressão ou classificação) e um dataset.
	Então	São apresentadas as informações: **do dataset**, com as respectivas features, e também **dos experimentos** a serem realizados, tais como profundidade máxima dos nós e funções a nível dos nós, dentre outras.

Tela(s) associadas ao [HU1] e **Cenário BDD** acima:

Command Window

```
GPTIPS 2 Demo 4: feature selection with concrete compressive strength data set
--------------------------------------------------------------------------------
The output being modelled is concrete compressive strength (MPa) and
the input variables are:

   Cement (x1) - kg in a m3 mixture
   Blast furnace slag (x2) - kg in a m3 mixture
   Fly ash (x3) - kg in a m3 mixture
   Water (x4) - kg in a m3 mixture
   Superplasticiser (x5) - kg in a m3 mixture
   Coarse aggregate (x6) - kg in a m3 mixture
   Fine aggregate (x7) - kg in a m3 mixture
   Age (x8) - range 1 - 365 days

To demonstrate feature selection in GPTIPS another 50 variables
consisting of normally distributed noise have been added to form the
input variables x9 to x58.

The configuration file is gpdemo4_config.m and the raw data is in
concrete.mat

The data has been divided into a training set, a holdout validation set
and a testing set.

GPTIPS is run twice for a maximum of 30 seconds per run or until a
RMSE of 6.5 is reached. The runs are merged into a single population
at the end.
```

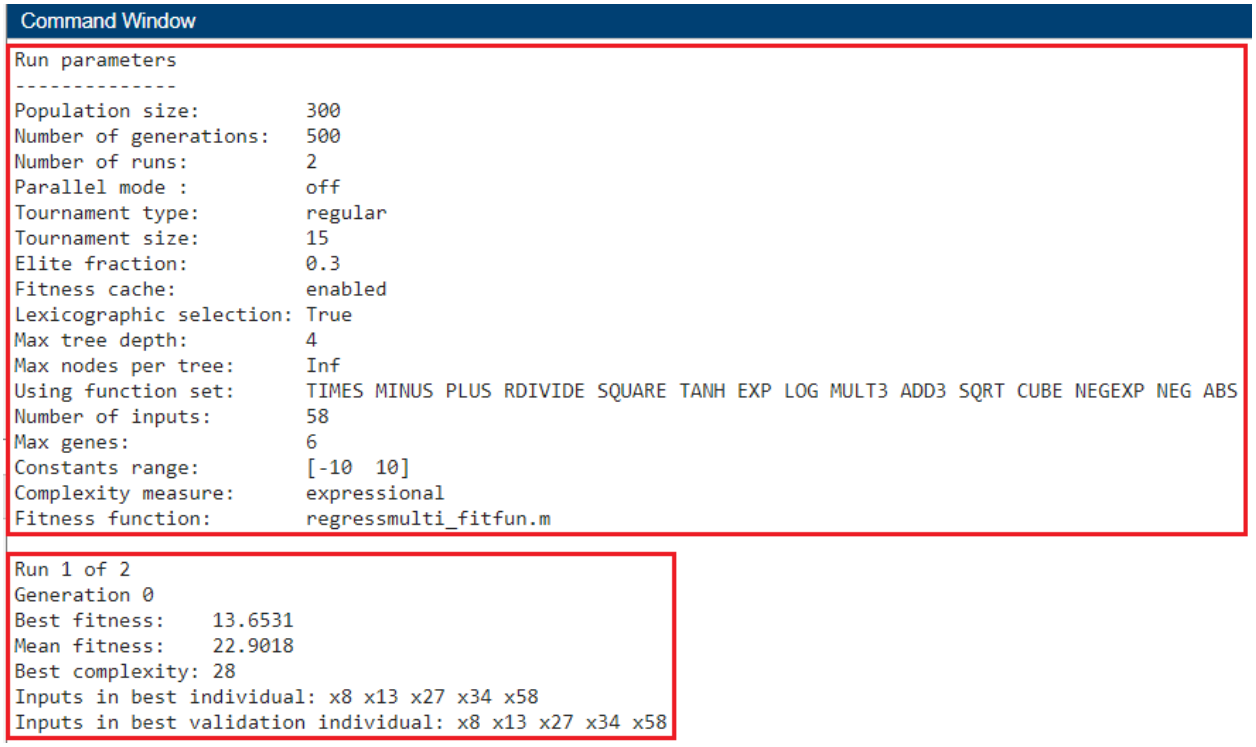
Command Window

```
6 genes are used (plus a bias term) so the form of the model will be
ypred = c0 + c1*tree1 + ... + c6*tree6
where ypred = predicted output, c0 = bias and c1,...,c6 are the gene weights.

Genes are limited to a depth of 4.

The function nodes used are:
TIMES MINUS PLUS RDIVIDE SQUARE TANH EXP LOG MULT3 ADD3 SQRT CUBE
POWER NEGEXP NEG ABS

The input variables that appear in the best model on the training
and validation data sets can be displayed at run time by
including the following two settings in gpdemo4_config.m :

gp.runcontrol.showBestInputs = true;
gp.runcontrol.showValBestInputs = true;

GPTIPS is run with the configuration in gpdemo4_config.m using :
>>gp=rungp(@gpdemo4_config);
Press a key to continue
```

Funcionalidade		
Como	usuário	

Quero	treinar modelos híbridos de IA + Econometria	
Para	tarefas de regressão ou classificação em datasets de dados seccionais ou de séries de tempo	
Cenário		
	Dado que	Selecionei, no script de seleção de tarefas e datasets, a tarefa e dataset de meu interesse.
	Quando	Decido por fazer o experimento.
	Então	São apresentados os parâmetros do experimento em sua integralidade, além das informações (por geração) de métrica do melhor indivíduo e métrica média da geração, melhor/menor complexidade apresentada, inputs/features no melhor indivíduo no conjunto de treino, inputs/features no melhor indivíduo no conjunto de validação.

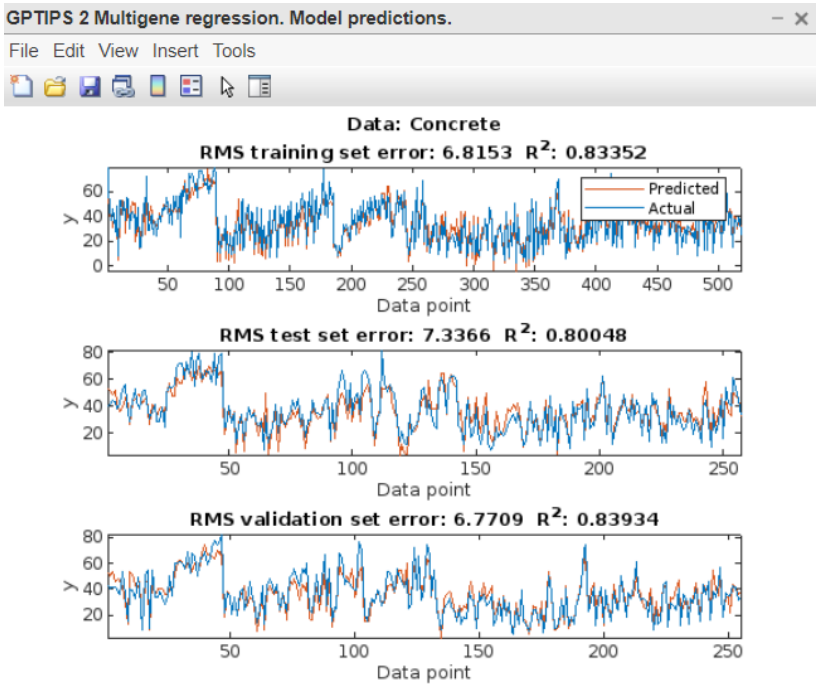
Tela(s) associadas ao [HU1] e **Cenário BDD** acima:

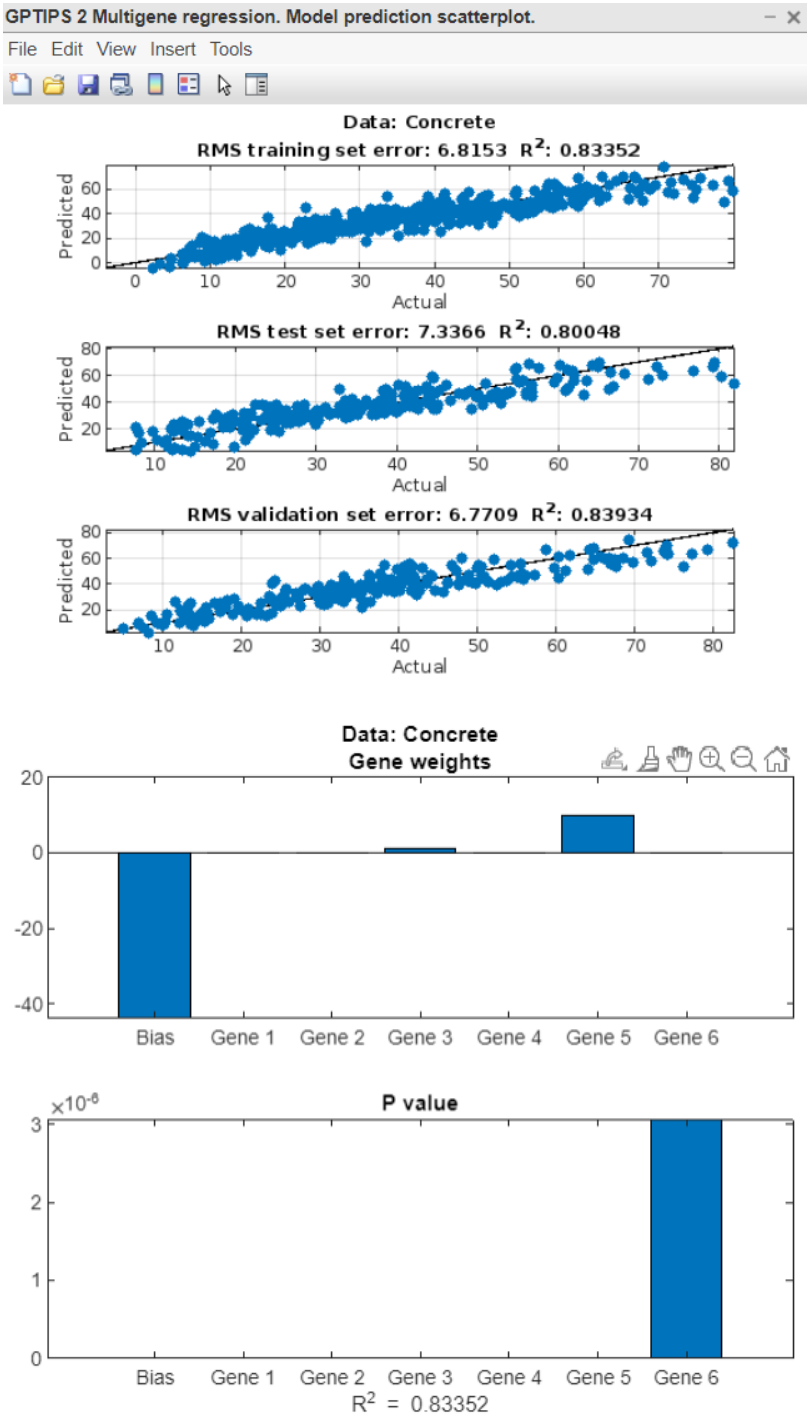


[HU2] Como [usuário] eu quero [ter acesso aos melhores modelos gerados pelo software, em função das métricas de acurácia disponíveis] para [avaliar se estes são competitivos frente aos SOTA].

Funcionalidade		
Como	usuário	
Quero	ter acesso aos melhores modelos gerados pelo software, em função das métricas de acurácia disponíveis	
Para	avaliar se estes são competitivos frente aos SOTA	
Cenário		
	Dado que	Selecionei, no script de seleção de tarefas e datasets, a tarefa e dataset de meu interesse.
	Quando	Decido por fazer o experimento.
	Então	São apresentados as métricas dos melhores modelos gerados pelo software, em função do RMSE e R2 Ajustado.

Tela(s) associadas ao [HU2] e **Cenário BDD** acima:




GPTIPS 2 Multigene regression. Model prediction scatterplot.
File Edit View Insert Tools
Data: Concrete
RMS training set error: 6.8153 R2: 0.83352
Predicted
Actual
RMS test set error: 7.3366 R2: 0.80048
Predicted
Actual
RMS validation set error: 6.7709 R2: 0.83934
Predicted
Actual
Data: Concrete
Gene weights
Bias
Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene 6
P value
x10-6
Bias
Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene 6
R2 = 0.83352

[HU3] Como [usuário] eu quero [averiguar a acurácia de um dado modelo gerado] para [avaliar se ele é suficiente para a tarefa real de regressão ou classificação que tenho, que exige um modelo com um mínimo de acurácia em casos reais].

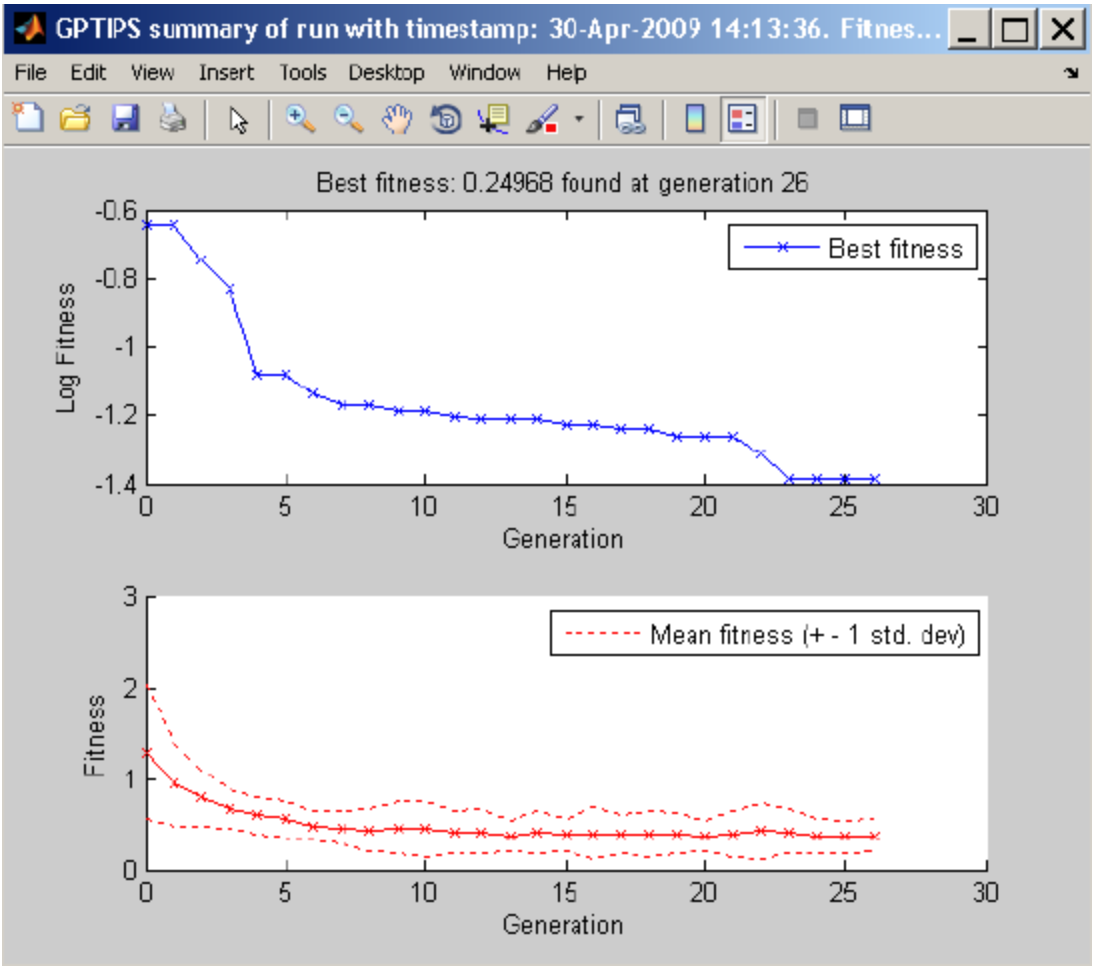
À [HU3] não foi representado qualquer Cenário BDD, pelo fato das telas e Cenários BDD feitos para a [HU2] traduzirem também a [HU3].

1.4.3. Casos de Uso (idealmente acompanhados de protótipos de telas).

Os Casos de Uso serão abreviados da seguinte forma: [CaUn].

Nome	[CaU1] Avaliar o Histórico da Métrica de Acurácia ao longo dos Experimentos, em especial das Gerações
Objetivo	Acompanhar o histórico da métrica de acurácia dos melhores indivíduos e médias de indivíduos ao longo da evolução, para que o usuário perceba se há melhora das populações.
Requisitos	RF1, RF2, RF3
Atores	Usuário
Prioridade	Média para Alta
Frequência de Uso	Alta
Criticidade	Média para Alta
Trigger	Ao realizar experimento
Fluxo Principal	1 - O usuário decide avaliar o Histórico **após** os experimentos. 2 - O usuário faz download dos dados do melhor indivíduo, de acurácia ao longo do tempo. 3 - O usuário exporta os resultados para relatório editor de texto.
Fluxo Alternativo	[A1] - O usuário decide avaliar o Histórico **durante** os experimentos. 1 - O software não permite que o usuário faça download dos dados do melhor indivíduo, de acurácia ao longo do tempo. 2 - O software pergunta, geração à geração, se o usuário deseja interromper os experimentos. Possível razão: após um dado número de gerações, os experimentos não mostram uma acurácia de melhor indivíduo em valor que seja razoável, no conhecimento tático do usuário. [A2] - O usuário decide por **não avaliar** o Histórico. 1 - O software executa em uma taxa de 20 a 25% melhor do que as opções anteriores. 2 - O software não mostra as informações do melhor indivíduo (acurácia ao longo do tempo).

Tela(s) associadas ao [CaU1]:
(as imagens postadas logo abaixo foram retiradas do **manual original do GPTIPS2**, pelo fato de ainda não estarem implementadas na versão **atualizada** do software que estou trabalhando)



Nome	[CaU2] Avaliar e Visualizar a Forma Simplificada de um Indivíduo.
Objetivo	Indivíduos gerados pelo software têm versões **não-simplificadas** (genes possuem expressões matemáticas das features originais, como a evolução possibilitou) ou **simplificadas** (simplificação matemática após terem sido combinadas, possivelmente 'somadas', todas as expressões matemáticas dentro dos genes). O objetivo é prover ao usuário uma versão simplificada, provavelmente menor, do indivíduo de interesse.
Requisitos	RF1, RF2, RF3
Atores	Usuário
Prioridade	Média para Alta
Frequência de Uso	Alta
Criticidade	Média para Alta

Trigger	Ao fim do experimento
Fluxo Principal	1 - O usuário opta por visualizar a forma simplificada de um indivíduo. 2 - Software avalia se a extensão GP.PRETTY está instalada, pois é ela que permite a visualização. 3 - Usuário visualiza o indivíduo na forma simplificada, e decide se armazenará o indivíduo nesta forma (simplificada) ou não-simplificada (original).
Fluxo Alternativo	[A1] - A extensão GP.PRETTY **não** está instalada. 1 - O usuário não conseguirá visualizar o indivíduo na forma simplificada. 2 - O software necessariamente armazenará o indivíduo na forma não-simplificada (original). [A2] - O usuário decide por não visualizar o indivíduo na forma simplificada. 1 - O software perguntará ao usuário, mesmo que ele(a) **não** deseje **visualizar** o indivíduo na forma simplificada, se o usuário deseja **armazenar** o indivíduo na forma simplificada. 2 - O software informa ao indivíduo o sucesso ou insucesso em armazenar o indivíduo.

Tela(s) associadas ao [CaU1]:
(as imagens postadas logo abaixo foram retiradas de **[4]**, pelo fato de ainda não estarem implementadas na versão **atualizada** do software que estou trabalhando)

$$y = \beta_1 x_1 + \beta_2 x_5 + \beta_3 x_6 + \beta_4 x_7 + \beta_5 x_9 + \beta_6 x_{11} + \beta_7 x_6 x_9 + \beta_8 x_6 x_{13} + \beta_9 x_9 x_{13} + \beta_{10} x_{13}^2 + \beta_{11} x_8 x_9 x_{11}^2 + \beta_{12} x_3 x_9 x_{10} + \beta_{13} x_3 x_{10} x_{13} + \beta_{14} x_8 x_9 x_{11} + \beta_{15} x_9 x_{10} x_{11} + \beta_{16} x_{10} x_{11} x_{13} + u$$

$$y = \beta_1 x_1 + \beta_2 x_4 + \beta_3 x_5 + \beta_4 x_6 + \beta_5 x_{11} + \beta_6 x_{12} + u$$

2. Projeto

A estrutura do projeto é organizada e mostrada em figuras, dispostas **do ponto de vista mais amplo ao mais específico**. O projeto é composto por arquivos *.m* (executáveis, caixa azuis) e *.mat* (bases de dados, caixas amarelas), algo potencialmente modificável, mediante interesse do usuário. Caixas cinzas devem ser desconsideradas.

Diagrama de Análise de Dependências em *view* Horizontal:

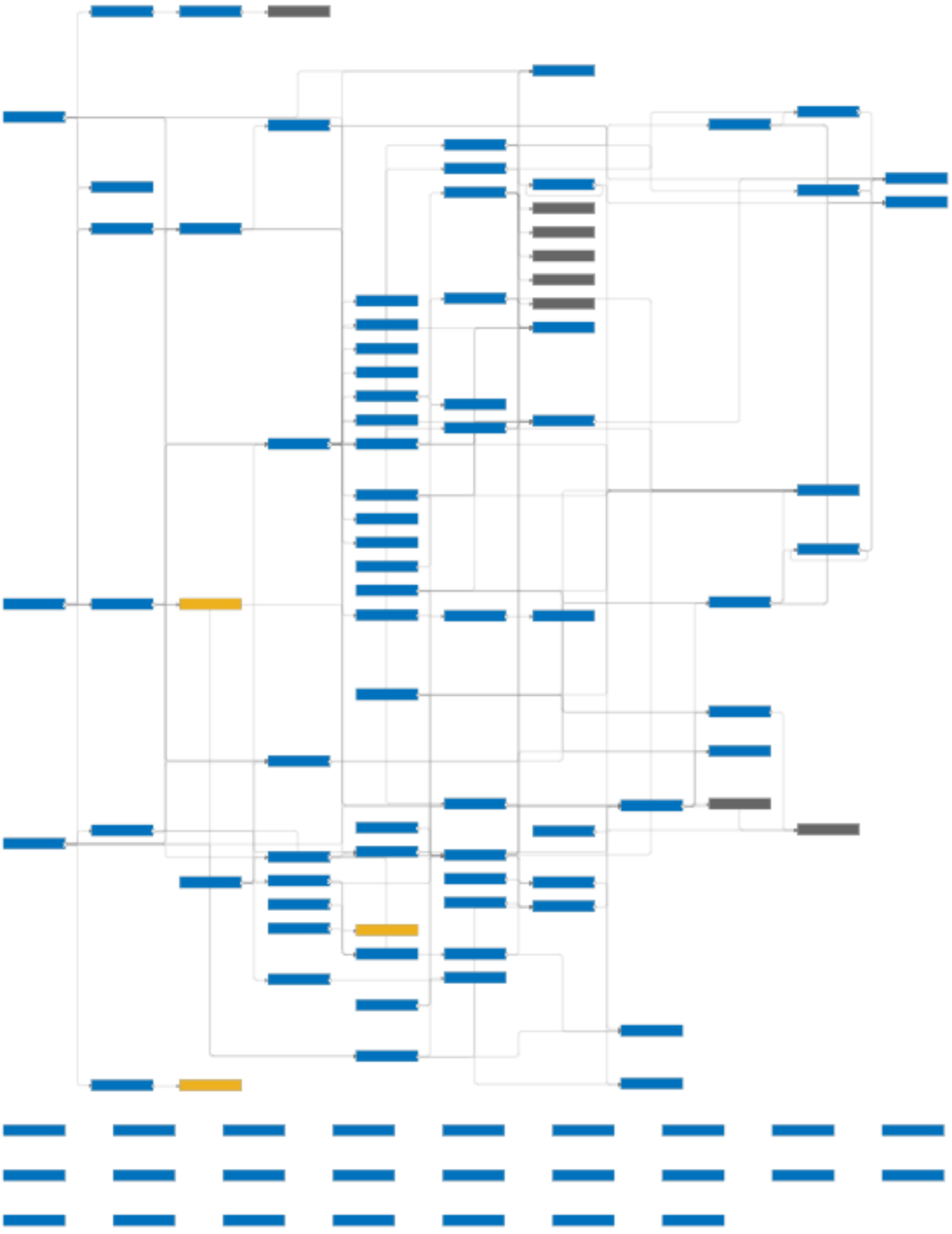
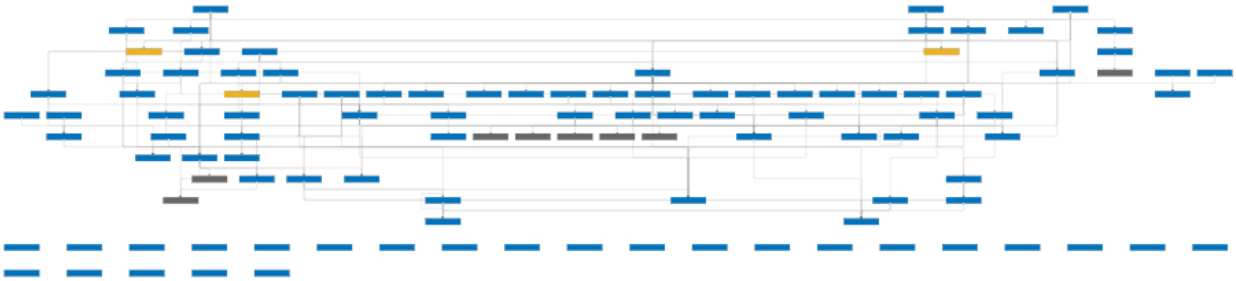
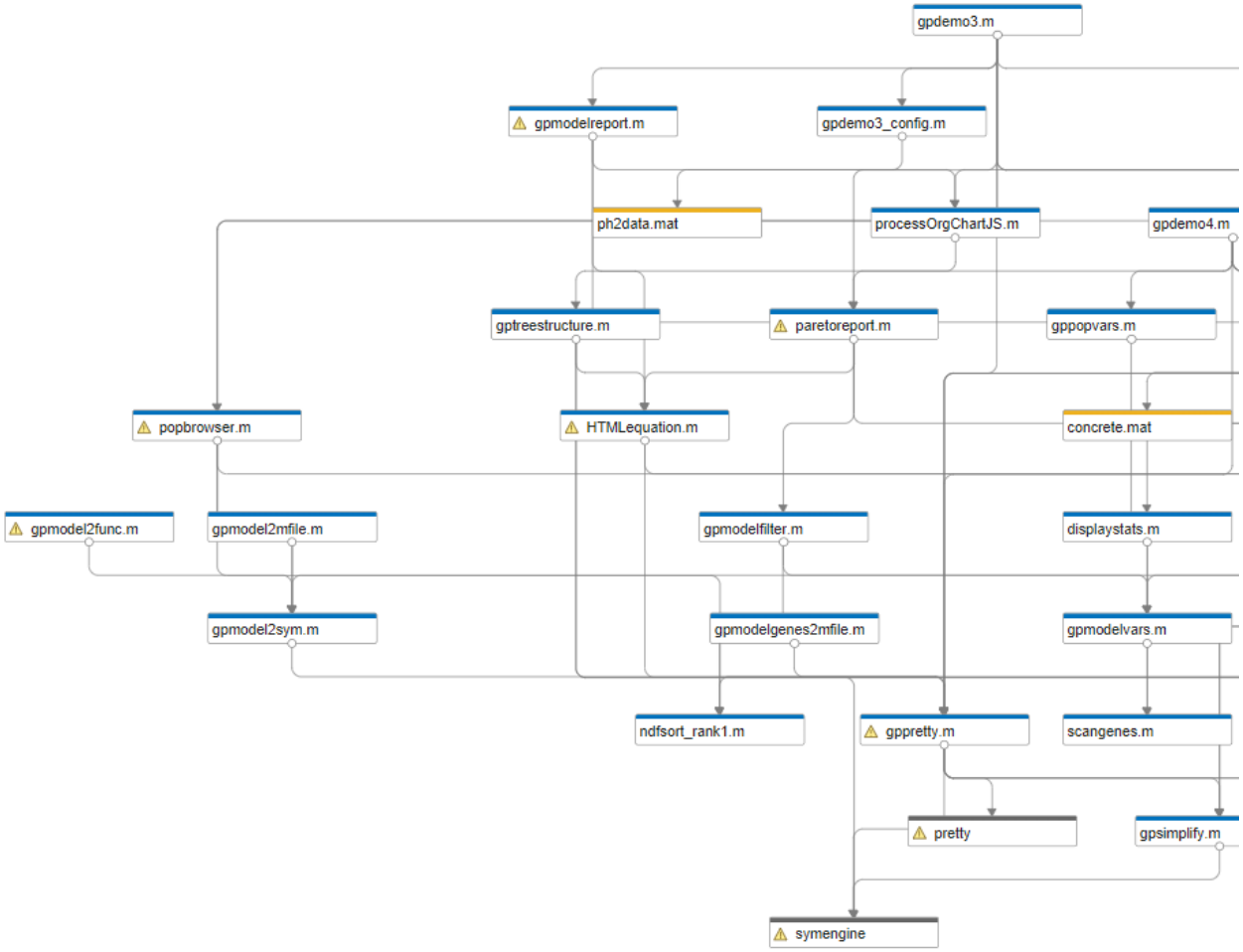
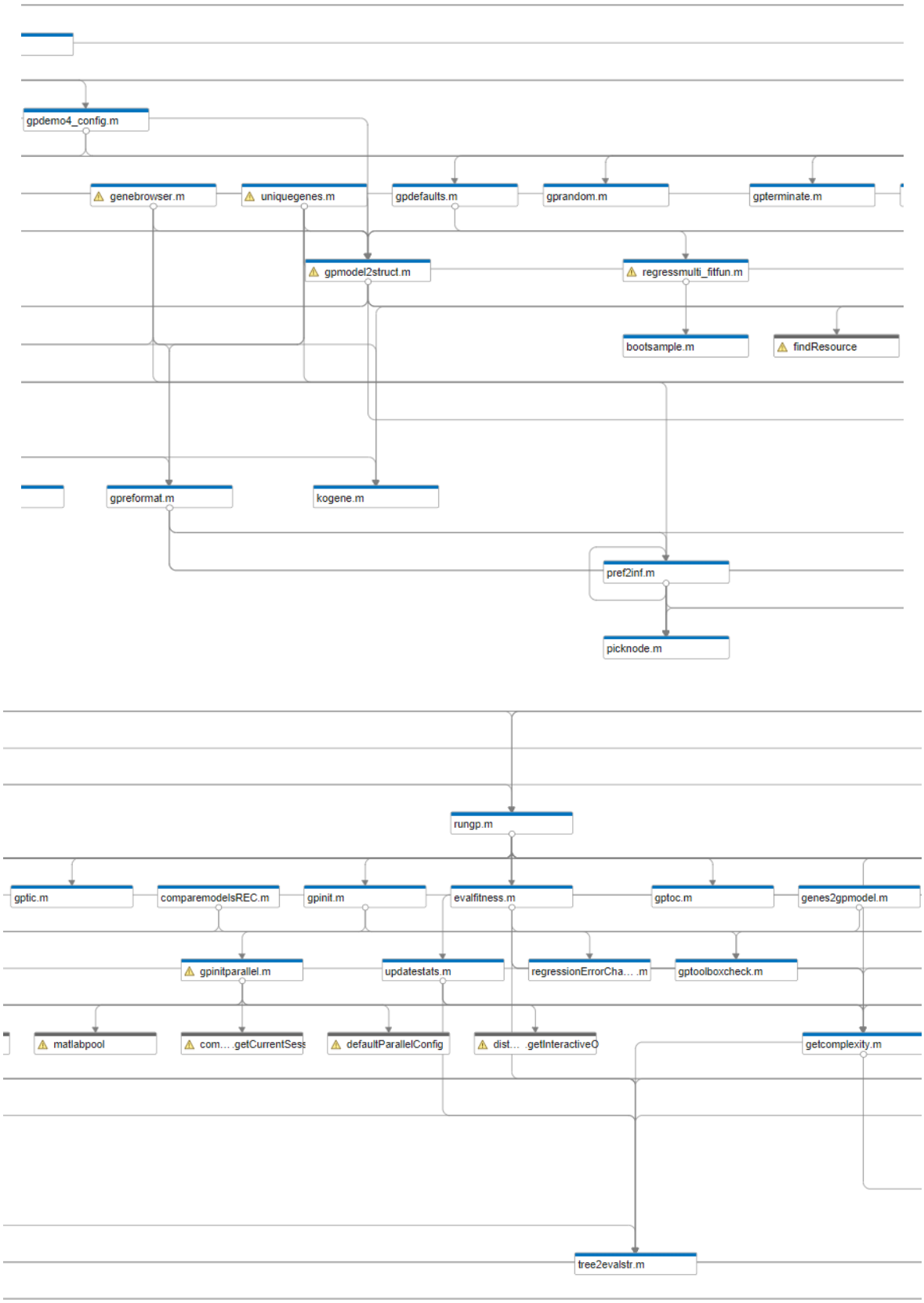


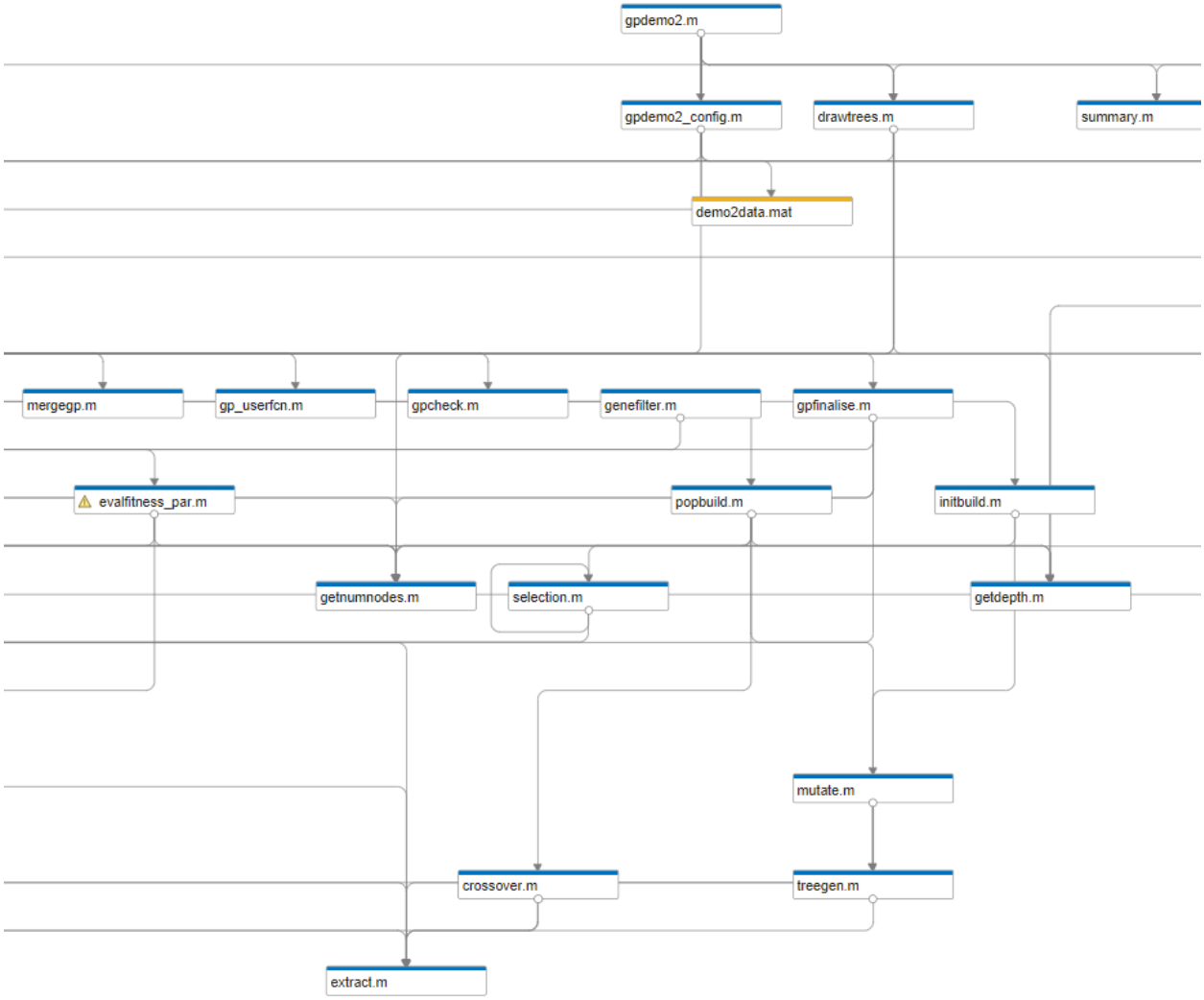
Diagrama de Análise de Dependências em *view* Vertical:

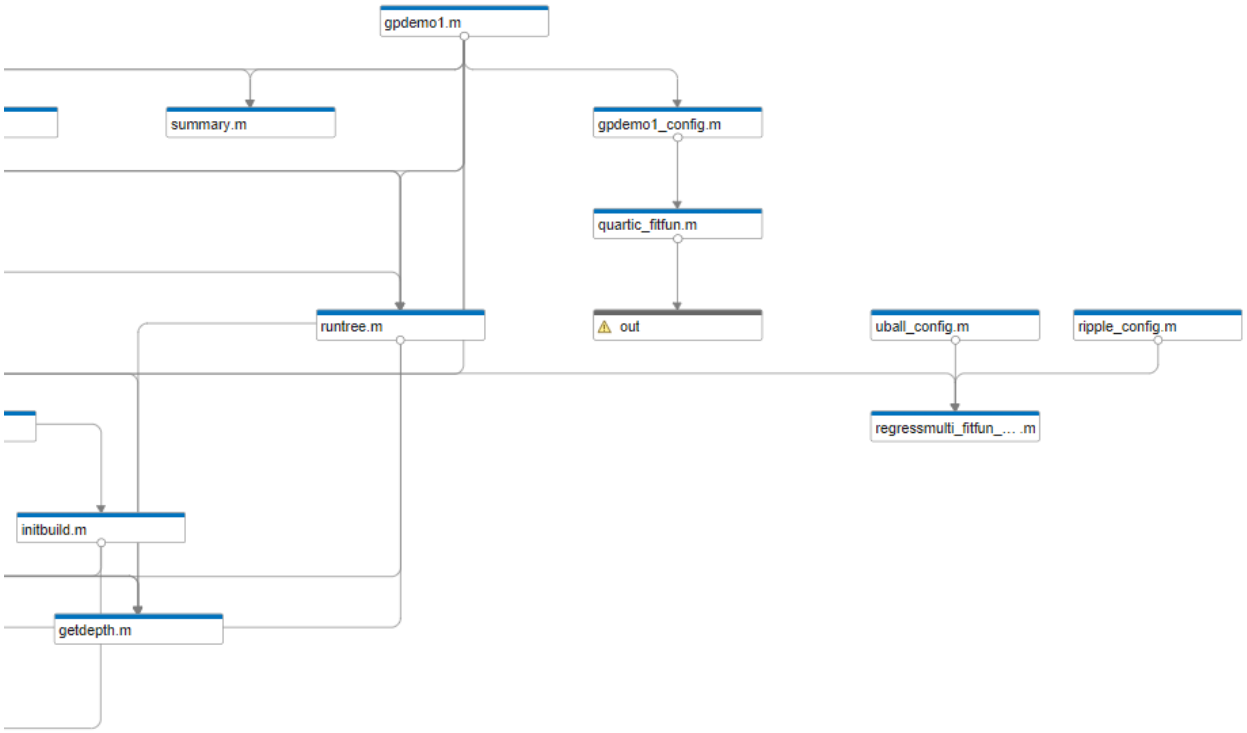


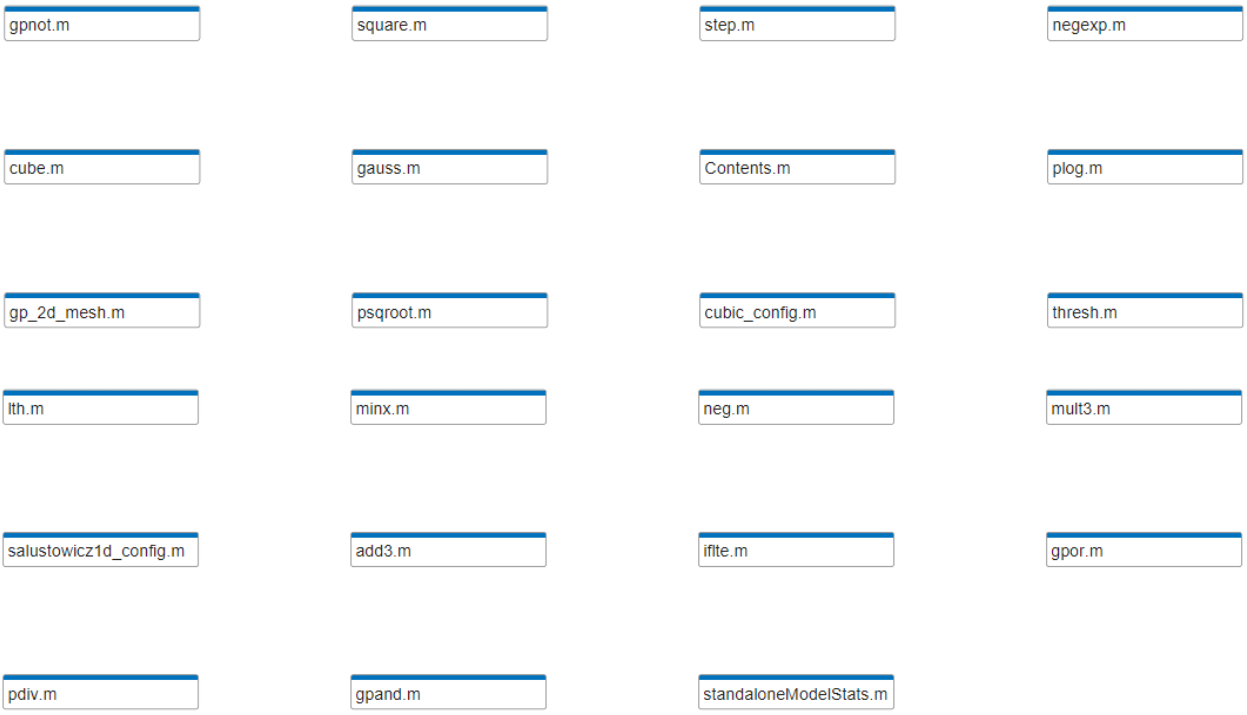
Zoom no **Diagrama de Análise de Dependências** em *view* Vertical:

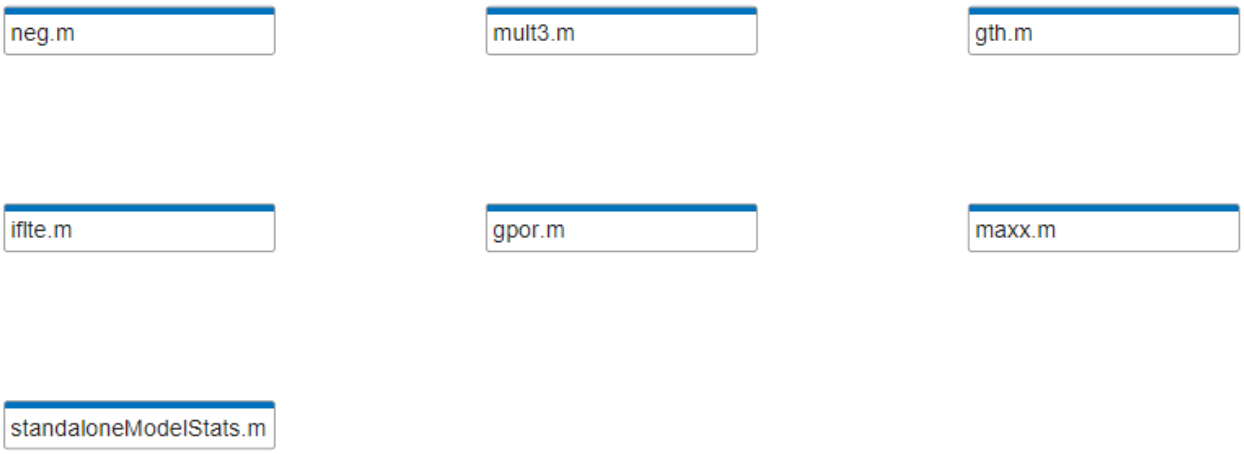



gpdemo4_config.m
genebrowser.m
uniquegenes.m
gpdefaults.m
gprandom.m
gpterminate.m
gpmodel2struct.m
regressmulti_fitfun.m
bootsample.m
findResource
gpreformat.m
kogene.m
pref2inf.m
picknode.m
rungp.m
gptic.m
comparemodelsREC.m
gpinit.m
evalfitness.m
gptoc.m
genes2gpmodel.m
gpinitparallel.m
updatestats.m
regressionErrorCha... .m
gptoolboxcheck.m
matlabpool
com....getCurrentSess
defaultParallelConfig
dist... .getInteractiveC
getcomplexity.m
tree2evalstr.m

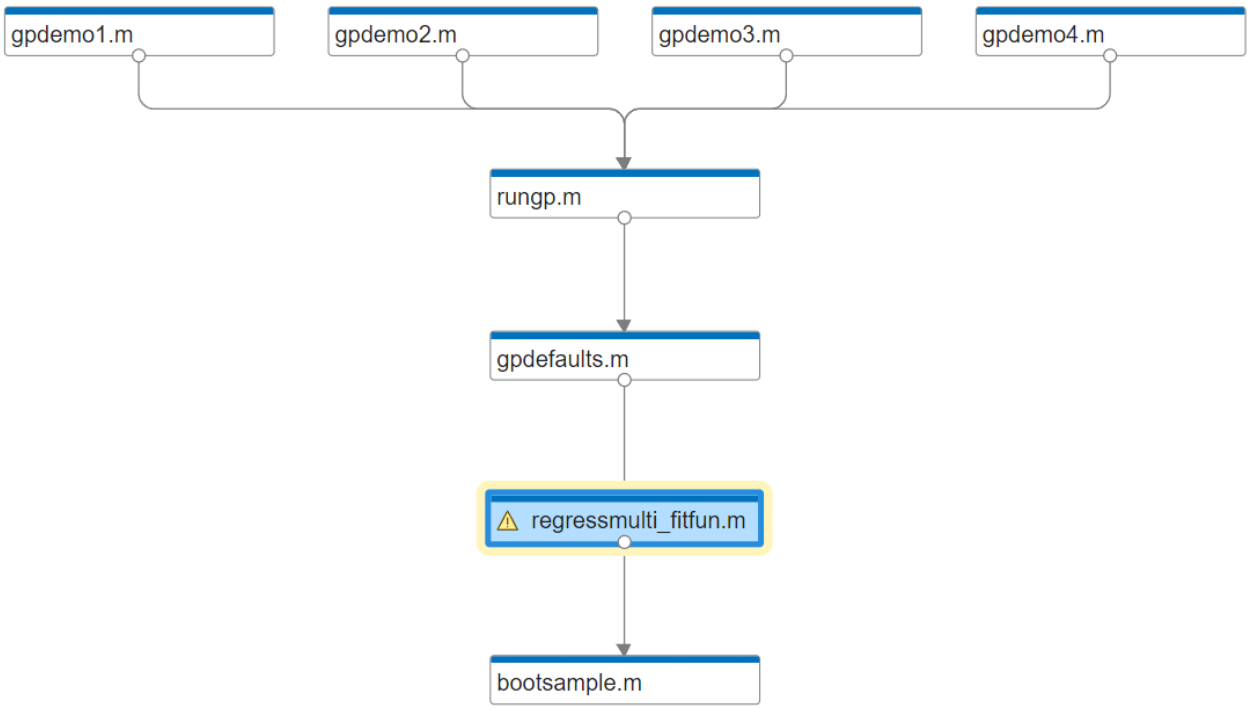

gpdemo2.m
gpdemo2_config.m
drawtrees.m
summary.m
demo2data.mat
mergegp.m
gp_userfcn.m
gpcheck.m
genefilter.m
gpfinalise.m
evalfitness_par.m
popbuild.m
initbuild.m
getnumnodes.m
selection.m
getdepth.m
mutate.m
crossover.m
treegen.m
extract.m


gpdemo1.m
summary.m
gpdemo1_config.m
quartic_fitfun.m
runtree.m
out
uball_config.m
ripple_config.m
regressmulti_fitfun_... .m
initbuild.m
getdepth.m


gpnot.m
square.m
step.m
negexp.m
cube.m
gauss.m
Contents.m
plog.m
gp_2d_mesh.m
psqroot.m
cubic_config.m
thresh.m
lth.m
minx.m
neg.m
mult3.m
salustowicz1d_config.m
add3.m
iflte.m
gpor.m
pdiv.m
gpand.m
standaloneModelStats.m



Abaixo, o **Diagrama de Análise de Dependências** em *view* Vertical para o script/função **regressmulti_fitfun**, responsável pela avaliação de um indivíduo.



3. Código Fonte

No repositório Git.

4. Testes

No repositório Git.

5. Documentação para o usuário

No repositório Git.

Datasets: são de escolha do usuário. Pré-carregados no programa estão os datasets de dados puramente seccionais *Resistência à Compressão do Concreto*, *Casas*, *Ruídos em Aerofólios*, *Propriedades Físico-Químicas da Estrutura Terciária de Proteínas*, *Estudo Hidrodinâmico de Iates*, *Câncer de Mama Wisconsin*, *Diabetes em Índios Pima*, e *Ionosfera*, datasets disponíveis no *UCI Machine Learning Repository*. Serão também carregados datasets de séries temporais.

Referências:

[1] Novaes, A. L. F., Tanscheit, R., & Dias, D. M. (2017, July). Econometric genetic programming outperforms traditional econometric algorithms for regression tasks. In Proceedings of the Genetic and Evolutionary Computation Conference Companion (pp. 1427-1430).

[2] Novaes, A. L. F., Tanscheit, R., & Dias, D. M. (2017, September). Econometric genetic programming in binary classification: evolving logistic regressions through genetic programming. In EPIA Conference on Artificial Intelligence (pp. 382-394). Springer, Cham.

[3] Novaes, A. L. F., Tanscheit, R., & Dias, D. M. (2016). Programação Genética Econométrica Aplicada a Problemas de Regressão em Conjuntos de Dados Seccionais. Proceedings of XIII Encontro Nacional de Inteligência Artificial, ENIAC.

[4] Novaes, A. L. F. (2015). Programação Genética Econométrica: uma Nova Abordagem para Problemas de Regressão e Classificação em Conjuntos de Dados Seccionais (Doctoral dissertation, Master's thesis. Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil).

[5] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting, 36(1), 54-74.

Siglas e Abreviações:

Inteligência Artificial (IA) e Deep Machine Learning (DML).