Data mrining

1. ການເຮັດ Apriori ຈະມີຈັກຂັ້ນຕອນ

   ມີ 2 ຂັ້ນຕອນ

2. ປະເພດຂອງຂໍ້ມູນ ມີຈັກປະເພດ
   ປະເພດຂອງຂໍ້ມູນໃນເວລາທີ່ ແບ່ງຕາມລັກສະນະຂອງຂໍ້ມູນ, ມີ 2 ປະເພດຂໍ້ມູນ, ຂໍ້ມູນປະລິມານ. ແລະຂໍ້ມູນຄຸນນະພາບ
3. Categorical Data ປະກອບມີ
   ແມ່ນຂໍ້ມູນທີ່ຖືກຈັດເປັນກຸ່ມເຊັ່ນ: ເພດ (ຊາຍ, ຍິງ), ການຈັດອັນດັບຂອງຂໍ້ມູນບໍ່ມີຄວາມຫມາຍ.
4. LHS (Left Hand Side) ສະແດງຮູບແບບ Itemset ທາງເບື້ອງຊ້າຍ ແມ່ນຄ່າ
   ບ່ອນທີ່ LHS (ຊ້າຍມື) ເປັນຕົວແທນຂອງຮູບແບບຂອງລາຍການທີ່ກໍານົດໄວ້ຢູ່ເບື້ອງຊ້າຍຂອງກົດລະບຽບການພົວພັນ, ແລະ RHS (ດ້ານຂວາມື) ເປັນຕົວແທນຂອງຮູບແບບຂອງລາຍການທີ່ກໍານົດໄວ້ຢູ່ເບື້ອງຂວາຂອງກົດລະບຽບການພົວພັນ.
5. ການຈັບກຸ່ມແບບ Hierarchicl Clustering ສາມາດເຮັດໄດ້ຈັກລັກສະນະ
   ການຈັບກຸ່ມແບບ Hierarchicl Clustering ສາມາດເຮັດໄດ້ມີ 5 ລັກສະນະ.

6. ຄໍາສັ່ງທີ່ໃຊ້ ໃນຕາຕະລາງນີ້ທັງໝົດມີຈັກຄັນ

7. ການປະມານຄ່າແບບຈຸດປະກອບມີ

8. DAD ຫຍໍ້ມາຈາກ

DAD: Democracy Assistance Dialogue.

9. ຂັ້ນຕອນນີ້ເປັນຂັ້ນຕອນທີ່ປ່ຽນແປງຂໍ້ມູນທີ່ໄດ້ເກັບລວບລວມມາ (rawdata) ໃຫ້ກາຍເປັນຂໍ້ ມູນທີ່ສາມາດນໍາໄປວິເຄາະໃນຂັ້ນຕອນຕໍ່ໄປໄດ້ ການແປງຂໍ້ມູນນີ້ອາດຈະຕ້ອງມີການ

ການກະກຽມຂໍ້ມູນແມ່ນຂັ້ນຕອນໃນຂະບວນການເຮັດວຽກ CRISP-DM ທີ່ປ່ຽນຂໍ້ມູນດິບໄປສູ່ຮູບແບບທີ່ສາມາດນໍາໃຊ້ໃນການວິເຄາະຕື່ມອີກ. ຂັ້ນຕອນນີ້ປະກອບມີວຽກງານເຊັ່ນ: ການທໍາຄວາມສະອາດແລະການຫັນປ່ຽນຂໍ້ມູນ, ການລຶບການຊ້ໍາກັນ, ການຈັດການຄ່າທີ່ຂາດຫາຍໄປ, ແລະການຈັດການກັບ outliers. ການກະກຽມຂໍ້ມູນແມ່ນສໍາຄັນຕໍ່ຄຸນນະພາບຂອງຜົນໄດ້ຮັບຈາກການຂຸດຄົ້ນຂໍ້ມູນ, ດັ່ງນັ້ນມັນເປັນສິ່ງສໍາຄັນເພື່ອຮັບປະກັນວ່າຂໍ້ມູນແມ່ນຖືກຕ້ອງ, ຄົບຖ້ວນສົມບູນ, ແລະກຽມພ້ອມສໍາລັບການວິເຄາະ. ຂໍ້ມູນທີ່ຖືກກະກຽມສາມາດຖືກນໍາໃຊ້ໃນຂັ້ນຕອນຕໍ່ໄປຂອງຂະບວນການຂຸດຄົ້ນຂໍ້ ມູນ.

10. ຄໍາສັ່ງຮັບຄ່າຈາກແປ້ນພິມໃນ Python

ການໄດ້ຮັບຄ່າຈາກແປ້ນພິມເປັນຮູບແບບການໂຕ້ຕອບຂອງຜູ້ໃຊ້. Python ສາມາດໄດ້ຮັບຄ່າໂດຍການໃຊ້ຟັງຊັນ input().

11. ຄໍາສັ່ງສະແດງຜົນບັງຄັບໃຫ້ຫລັງຈຸດເປັນ 2 ຕົວ

12. ໃນຂະບວນການເຮັດວຽກຂອງ CRISP-DM ຜົນໄດ້ຮັບທີ່ໄດ້ຈາກການວິເຄາະດ້ວຍເທັກນິກ ການຂຸດຄົ້ນຂໍ້ມູນ ເຖິງແມ່ນວ່າຜົນທີ່ໄດ້ຮັບ ທີ່ສະແດງເຖິງປະໂຫຍດຄວາມຮູ້ ແລະ ນໍາເອົາ ຄວາມຮູ້ເຂົ້າ ນີ້ໄປນໍາໃຊ້ຕົວຈິງໃນອົງກອນ ຫລື ບໍລິສັດ

ໃນຂະບວນການເຮັດວຽກຂອງ CRISP-DM, ຜົນໄດ້ຮັບຂອງການວິເຄາະການຂຸດຄົ້ນຂໍ້ມູນແມ່ນຂັ້ນຕອນສຸດທ້າຍທີ່ສະແດງໃຫ້ເຫັນເຖິງຜົນປະໂຫຍດຂອງຄວາມຮູ້ທີ່ມາຈາກຂະບວນການຂຸດຄົ້ນຂໍ້ມູນ. ຂັ້ນຕອນນີ້ປະກອບມີການປະເມີນຜົນໄດ້ຮັບ, ພິຈາລະນາຜົນກະທົບຂອງມັນ, ແລະກໍານົດວິທີການທີ່ພວກມັນຈະຖືກນໍາມາໃຊ້ໃນການຕັດສິນໃຈແລະຊຸກຍູ້ການປະຕິບັດພາຍໃນອົງກອນ. ຜົນໄດ້ຮັບຂອງການຂຸດຄົ້ນຂໍ້ມູນສາມາດມີຜົນກະທົບຢ່າງຫຼວງຫຼາຍ, ສະນັ້ນມັນເປັນສິ່ງສໍາຄັນທີ່ຈະສື່ສານຜົນໄດ້ຮັບກັບຜູ້ທີ່ກ່ຽວຂ້ອງທີ່ກ່ຽວຂ້ອງແລະຮັບປະກັນວ່າພວກເຂົາຖືກນໍາໃຊ້ໃນທາງທີ່ມີຄວາມຫມາຍ.

13. ຄໍາສັ່ງທີ່ໃຊ້ ຕັດຖັນທີ່ມີຂໍ້ມູນເປັນ null ອອກ`dataset_df1= dataset.loc[:,(dataset.notnull().any())]`

14. ປະເພດຂອງການປະເມີນຄ່າປະກອບມີ

ມີຫຼາຍປະເພດຂອງການປະເມີນທີ່ສາມາດນໍາໃຊ້ເພື່ອປະເມີນການປະຕິບັດຂອງແບບຈໍາລອງການຂຸດຄົ້ນຂໍ້ມູນ, ລວມທັງ

- ການປະເມີນຄວາມຖືກຕ້ອງ,
- ຄວາມຊັດເຈນແລະການປະເມີນຄືນ,
- ການປະເມີນເສັ້ນໂຄ້ງ ROC,
- ການປະເມີນຕາຕະລາງຍິກ,
- ການປະເມີນ F-Measure,
- ການປະເມີນຄວາມຖືກຕ້ອງຂ້າມ
- ການປະເມີນແບບຈໍາລອງ.

15. ຈົ່ງໃຫ້ຄວາມໝາຍ, ຂະບວນການ ແລະ ປະໂຫຍດ ຂອງການຂຸດຄົ້ນຂໍ້ມູນ?

ການຂຸດຄົ້ນຂໍ້ມູນແມ່ນຂະບວນການຄົ້ນພົບຮູບແບບແລະຄວາມຮູ້ຈາກຂໍ້ມູນຈໍານວນຫລາຍໂດຍໃຊ້ເຕັກນິກສະຖິຕິ, ຂັ້ນຕອນການຮຽນຮູ້ເຄື່ອງຈັກ, ແລະເຕັກໂນໂລຢີຖານຂໍ້ມູນ. ເປົ້າຫມາຍແມ່ນເພື່ອສະກັດຄວາມເຂົ້າໃຈທີ່ສາມາດນໍາໃຊ້ໃນການຕັດສິນໃຈທີ່ມີຂໍ້ມູນແລະປັບປຸງຂະບວນການທຸລະກິດ. ຂະບວນການປະກອບມີການກະກຽມຂໍ້ມູນ, ການຂຸດຄົ້ນ, ການຄັດເລືອກຕົວແບບ, ການກໍ່ສ້າງ, ແລະການນໍາໃຊ້. ການຂຸດຄົ້ນຂໍ້ມູນສາມາດນໍາເອົາຜົນປະໂຫຍດທີ່ສໍາຄັນເຊັ່ນ: ການປັບປຸງການຕັດສິນໃຈ, ການເພີ່ມປະສິດທິພາບ, ການແຂ່ງຂັນທີ່ເພີ່ມຂຶ້ນ, ແລະຄວາມພໍໃຈຂອງລູກຄ້າທີ່ດີຂຶ້ນ. ຢ່າງໃດກໍ່ຕາມ, ມັນເປັນສິ່ງສໍາຄັນທີ່ຈະພິຈາລະນາຂໍ້ຈໍາກັດເຊັ່ນຄວາມລໍາອຽງ, ຄວາມຕ້ອງການຄວາມຊໍານານພິເສດ, ແລະສິ່ງທ້າທາຍຂອງການຮັບປະກັນຄວາມເປັນສ່ວນຕົວແລະຄວາມປອດໄພຂອງຂໍ້ມູນ.

16. ຮູບແບບການຂຸດຄົ້ນຂໍ້ມູນມີຈັກຮູບແບບ ໃຫ້ອະທິບາຍ ແລະ ຍົກຕົວຢ່າງ ແຕ່ລະຮູບແບບ?

- Decision Trees: ປະເພດຂອງຮູບແບບນີ້ແມ່ນເປັນຕົວແທນຂອງໂຄງສ້າງຕົ້ນໄມ້, ບ່ອນທີ່ແຕ່ລະ node ເປັນຕົວແທນຂອງການຕັດສິນໃຈໂດຍອີງໃສ່ລັກສະນະສະເພາະໃດຫນຶ່ງ, ແລະ ສາຂາເປັນຕົວແທນຂອງຜົນໄດ້ຮັບທີ່ເປັນໄປໄດ້ຂອງການຕັດສິນໃຈ.
- Neural Networks: ປະເພດຂອງຕົວແບບນີ້ແມ່ນໄດ້ຮັບການດົນໃຈຈາກໂຄງສ້າງແລະຫນ້າທີ່ຂອງສະຫມອງຂອງມະນຸດແລະຖືກນໍາໃຊ້ສໍາລັບວຽກງານເຊັ່ນ: ການຈັດປະເພດຮູບພາບແລະ ການຮັບຮູ້ສຽງເວົ້າ.
- Support Vector Machines (SVMs):: ປະເພດຂອງແບບຈໍາລອງນີ້ແມ່ນໃຊ້ສໍາລັບບັນຫາການຈັດປະເພດແລະເຮັດວຽກໂດຍການຊອກຫາເຂດແດນທີ່ດີທີ່ສຸດທີ່ແຍກຊັ້ນໃນຂໍ້ມູນ. SVMs ຖືກນໍາໃຊ້ຢ່າງກວ້າງຂວາງສໍາລັບວຽກງານເຊັ່ນ: ການຈັດປະເພດຂໍ້ຄວາມແລະການຮັບຮູ້ຮູບພາບ.
- Clustering Models: ແບບຈໍາລອງປະເພດນີ້ແມ່ນໃຊ້ສໍາລັບການຈັດກຸ່ມຈຸດຂໍ້ມູນທີ່ຄ້າຍຄືກັນຮ່ວມກັນ ແລະຖືກນໍາໃຊ້ທົ່ວໄປສໍາລັບການແບ່ງສ່ວນຕະຫຼາດ ແລະການສ້າງໂປຣໄຟລ໌ລູກຄ້າ.
- Association Rules : ຮູບແບບປະເພດນີ້ແມ່ນນໍາໃຊ້ສໍາລັບການວິເຄາະກະຕ່າຕະຫຼາດແລະການນໍາໃຊ້ເພື່ອຊອກຫາຄວາມສໍາພັນລະຫວ່າງລາຍການໃນຊຸດຂໍ້ມູນ. ສູດການຂຸດຄົ້ນຂອງກົດລະບຽບຂອງສະມາຄົມລວມມີ algorithm Apriori ແລະ algorithm ECLAT.
- Linear Regression: ແບບຈໍາລອງປະເພດນີ້ແມ່ນໃຊ້ເພື່ອຄາດຄະເນຕົວແປເປົ້າໝາຍຢ່າງຕໍ່ເນື່ອງໂດຍອີງໃສ່ຊຸດຄຸນສົມບັດການປ້ອນຂໍ້ມູນ.
- Logistic Regression: ຮູບແບບນີ້ຖືກນໍາໃຊ້ສໍາລັບບັນຫາການຈັດປະເພດຄູ່, ບ່ອນທີ່ເປົ້າຫມາຍແມ່ນເພື່ອຄາດຄະເນຫນຶ່ງໃນສອງຜົນໄດ້ຮັບທີ່ເປັນໄປໄດ້. Logistic regression ສ້າງຄວາມສໍາພັນລະຫວ່າງລັກສະນະການປ້ອນຂໍ້ມູນແລະຕົວແປເປົ້າຫມາຍເປັນຫນ້າທີ່ logistic.
- Naive Bayes: ແບບຈໍາລອງປະເພດນີ້ແມ່ນໃຊ້ສໍາລັບບັນຫາການຈັດປະເພດແລະເຮັດວຽກໂດຍໃຊ້ທິດສະດີຂອງ Bayes ເພື່ອຄິດໄລ່ຄວາມເປັນໄປໄດ້ຂອງຫ້ອງຮຽນທີ່ໃຫ້ຊຸດຂອງ ລັກສະນະການປ້ອນຂໍ້ມູນ. Naive Bayes ມັກຈະຖືກນໍາໃຊ້ສໍາລັບການຈັດປະເພດຂໍ້ຄວາມແລະການວິເຄາະຄວາມຮູ້ສຶກ.

17. KDD ຫຍໍ້ມາຈາກ

KDD: Knowledge Discovery in Database.

18. ການຈັດກຸ່ມຍັງສາມາດເຮັດໄດ້ຈັກລັກສະນະ

ມີຫຼາຍປະເພດຂອງການເຊື່ອມໂຍງທີ່ສາມາດເຮັດໄດ້, ລວມທັງ:

- ການເຊື່ອມໂຍງຂໍ້ມູນ: ນີ້ກ່ຽວຂ້ອງກັບການລວມເອົາຂໍ້ມູນຈາກຫຼາຍແຫຼ່ງເຂົ້າໄປໃນມຸມເບິ່ງລວມດຽວ, ເຊັ່ນ: ການລວມຂໍ້ມູນຈາກຖານຂໍ້ມູນທີ່ແຕກຕ່າງກັນຫຼືຄັງຂໍ້ມູນ.
- ການປະສົມປະສານ Algorithm: ນີ້ກ່ຽວຂ້ອງກັບການລວມເອົາລະບົບການຂຸດຄົ້ນຂໍ້ມູນຫຼາຍອັນເພື່ອປັບປຸງຜົນໄດ້ຮັບທີ່ໄດ້ຮັບຈາກສູດການຄິດໄລ່ດຽວ.
- ການເຊື່ອມໂຍງແບບຈໍາລອງ: ນີ້ກ່ຽວຂ້ອງກັບການລວມຕົວແບບຫຼາຍແບບ, ເຊັ່ນ: ການຕັດຕົ້ນໄມ້, ເຄືອຂ່າຍ neural, ຫຼືເຄື່ອງ vector ສະຫນັບສະຫນຸນ, ເພື່ອປັບປຸງຄວາມຖືກຕ້ອງຂອງ ການຄາດຄະເນ.
- ການເຊື່ອມໂຍງໂດເມນ: ນີ້ກ່ຽວຂ້ອງກັບການລວມເອົາຄວາມຮູ້ຈາກໂດເມນທີ່ແຕກຕ່າງກັນ, ເຊັ່ນ: ຂໍ້ມູນທາງການແພດ, ທາງດ້ານການເງິນ, ແລະສື່ສັງຄົມ, ເພື່ອສ້າງມຸມເບິ່ງທີ່ຄົບຖ້ວນແລະ ຖືກຕ້ອງຂອງບັນຫາສະເພາະ.
- ການປະສົມປະສານຄວາມຮູ້: ນີ້ກ່ຽວຂ້ອງກັບການລວມເອົາຄວາມຮູ້ຈາກແຫຼ່ງຕ່າງໆ, ເຊັ່ນຄວາມຮູ້ຂອງຜູ້ຊ່ຽວຊານ, ຂໍ້ມູນປະຫວັດສາດ, ແລະຂໍ້ມູນໃນເວລາທີ່ແທ້ຈິງ, ເພື່ອປັບປຸງຜົນໄດ້ຮັບທີ່ ໄດ້ຮັບຈາກການຂຸດຄົ້ນຂໍ້ມູນ.
- ການເຊື່ອມໂຍງລະຫວ່າງລະບຽບວິໄນ: ນີ້ກ່ຽວຂ້ອງກັບການລວມເອົາການຂຸດຄົ້ນຂໍ້ມູນກັບວິຊາອື່ນໆ, ເຊັ່ນການຮຽນຮູ້ເຄື່ອງຈັກ, ປັນຍາປະດິດ, ແລະສະຖິຕິ, ເພື່ອປັບປຸງຜົນໄດ້ຮັບທີ່ໄດ້ຮັບ ຈາກການຂຸດຄົ້ນຂໍ້ມູນ.

19. ເຕັກນິກການແບ່ງກຸ່ມຂໍ້ມູນຕາມຄຸນລັກສະນະຕ່າງໆ ທີ່ໄດ້ມີການກໍານົດໄວ້ເພື່ອສ້າງຕົວແບບ ສໍາລັບຮູບແບບການຄໍາດເດົາໃນອະນາຄົດ, ເອີ້ນວ່າ (Supervised learning)

ແມ່ນປະເພດຂອງການຮຽນຮູ້ຂອງເຄື່ອງຈັກໃນການຂຸດຄົ້ນຂໍ້ມູນທີ່ສຸດການຄິດໄລ່ໄດ້ຖືກຝຶກອົບຮົມກ່ຽວກັບຂໍ້ມູນທີ່ມີປ້າຍຊື່ເພື່ອຄາດຄະເນຕົວແປເປົ້າຫມາຍໂດຍອີງໃສ່ຊຸດຂອງລັກສະນະການປ້ອນຂໍ້ ມູນ. ໃນຄໍາສັບຕ່າງໆອື່ນໆ, ສູດການຄິດໄລ່ແມ່ນໃຫ້ຊຸດຂໍ້ມູນທີ່ມີປ້າຍຊື່, ແລະມັນໃຊ້ຂໍ້ມູນນີ້ເພື່ອຮຽນຮູ້ຄວາມສໍາພັນລະຫວ່າງລັກສະນະການປ້ອນຂໍ້ມູນແລະຕົວແປເປົ້າຫມາຍ. ເມື່ອຕົວແບບໄດ້ຮັບການ ຝຶກອົບຮົມ, ມັນສາມາດຖືກນໍາໃຊ້ເພື່ອເຮັດໃຫ້ການຄາດຄະເນກ່ຽວກັບຂໍ້ມູນໃຫມ່ທີ່ບໍ່ມີປ້າຍຊື່.

ບາງຕົວຢ່າງທົ່ວໄປຂອງການຮຽນຮູ້ທີ່ມີການເບິ່ງແຍງໃນການຂຸດຄົ້ນຂໍ້ມູນປະກອບມີ:

- Regression: ປະເພດຂອງການຮຽນຮູ້ການເບິ່ງແຍງນີ້ແມ່ນໃຊ້ເພື່ອຄາດຄະເນຕົວແປເປົ້າຫມາຍຢ່າງຕໍ່ເນື່ອງ, ເຊັ່ນລາຄາຂອງເຮືອນຫຼືຄວາມເປັນໄປໄດ້ຂອງ churning ລູກຄ້າ.
- ການຈັດປະເພດ: ປະເພດຂອງການຮຽນຮູ້ການເບິ່ງແຍງນີ້ແມ່ນໃຊ້ເພື່ອຄາດຄະເນຕົວແປເປົ້າຫມາຍປະເພດເຊັ່ນວ່າລູກຄ້າຈະຊື້ສິນຄ້າຫຼືບໍ່.
- ການກວດຫາຄວາມຜິດປົກກະຕິ: ປະເພດຂອງການຮຽນຮູ້ແບບຄວບຄຸມນີ້ຖືກນໍາໃຊ້ເພື່ອກໍານົດຈຸດຂໍ້ມູນທີ່ deviate ຈາກພຶດຕິກໍາປົກກະຕິໃນຊຸດຂໍ້ມູນ.

20. Market Basket Analysis ແມ່ນຫຍັງ?

ແມ່ນມັນເປັນຮູບແບບການນໍາໃຊ້ເພື່ອຊອກຫາກຸ່ມຂອງວັດຖຸທີ່ມີແນວໂນ້ມທີ່ຈະປະກົດຢູ່ຮ່ວມກັນໃນການເຮັດທຸລະກໍາ, ມັກຈະເປັນການເຮັດທຸລະກໍາຈຸດຂອງການຂາຍຜົນໄດ້ຮັບຍັງສາມາດ ສະແດງໄດ້. ກົດຫມາຍວ່າດ້ວຍການພົວພັນ, ທີ່ກໍານົດຄວາມເປັນໄປໄດ້ຂອງການຊື້ສິ່ງຂອງຮ່ວມກັນ.

21. ສາເຫດສ່ວນໃຫ່ຍຂອງຄ່າຜິດປົກກະຕິໃນຊຸດຂໍ້ມູນ:

ມີຫຼາຍເຫດຜົນວ່າ ອາດຈະປາກົດຢູ່ໃນຊຸດຂໍ້ມູນໃນການຂຸດຄົ້ນຂໍ້ມູນ, ລວມທັ້ງ:

- ຄວາມຜິດພາດໃນການປ້ອນຂໍ້ມູນ: ຄວາມຜິດພາດຂອງມະນຸດໃນລະຫວ່າງການປ້ອນຂໍ້ມູນສາມາດເຮັດໃຫ້ເກີດຄວາມຜິດປົກກະຕິເຊັ່ນ: ພິມຜິດ ຫຼືຄ່າທີ່ບໍ່ຖືກຕ້ອງຖືກປ້ອນ.
- ຄວາມຜິດພາດໃນການວັດແທກ: ຕົວເລກທີ່ເກີນອາດຈະເກີດຈາກຄວາມຜິດພາດການວັດແທກເຊັ່ນ: ເຄື່ອງມືທີ່ຜິດພາດ ຫຼືການປັບທຽບບໍ່ຖືກຕ້ອງ.
- ການປ່ຽນແປງທາງທໍາມະຊາດ: Outliers ຍັງສາມາດເກີດຂຶ້ນຕາມທໍາມະຊາດໃນຊຸດຂໍ້ມູນອັນເນື່ອງມາຈາກຄວາມແຕກຕ່າງກັນໃນຂໍ້ມູນ. ຕົວຢ່າງ, ໃນຊຸດຂໍ້ມູນຂອງການຊື້ຂອງລູກຄ້າ, ອາດຈະມີລູກຄ້າບາງຄົນທີ່ເຮັດການຊື້ຂະຫນາດໃຫຍ່ຫຼາຍກ່ວາລູກຄ້າສ່ວນໃຫຍ່.
- ຄວາມລໍາອຽງຂອງຕົວຢ່າງ: Outliers ຍັງສາມາດເປັນຜົນມາຈາກຄວາມລໍາອຽງຂອງຕົວຢ່າງ, ບ່ອນທີ່ຕົວຢ່າງທີ່ໃຊ້ໃນການສ້າງຊຸດຂໍ້ມູນບໍ່ແມ່ນຕົວແທນຂອງປະຊາກອນ.
- Concept Drift: Outliers ຍັງອາດຈະປາກົດຂຶ້ນເນື່ອງຈາກການປ່ຽນແປງການກະຈາຍຂໍ້ມູນພື້ນຖານໃນໄລຍະເວລາ, ເອີ້ນວ່າແນວຄວາມຄິດ drift.
- Data Corruption: ໃນບາງກໍລະນີ, outliers ອາດເປັນຜົນມາຈາກການສໍ້ລາດບັງຫຼວງຂອງຂໍ້ມູນ, ບ່ອນທີ່ມູນຄ່າຂໍ້ມູນໄດ້ຖືກປ່ຽນແປງຫຼື manipulated.

22. ສຸບແບບການຂຸດຄົ້ນຂໍ້ມູນ

- ຕົ້ນໄມ້ການຕັດສິນໃຈ: ປະເພດຂອງສຸບແບບນີ້ແມ່ນເປັນຕົວແທນຂອງໂຄງສ້າງຕົ້ນໄມ້, ບ່ອນທີ່ແຕ່ລະ node ເປັນຕົວແທນຂອງການຕັດສິນໃຈໂດຍອີງໃສ່ລັກສະນະສະເພາະໃດຫນຶ່ງ, ແລະ ສາຂາເປັນຕົວແທນຂອງຜົນໄດ້ຮັບທີ່ເປັນໄປໄດ້ຂອງການຕັດສິນໃຈ.
- ເຄືອຂ່າຍ neural: ປະເພດຂອງຕົວແບບນີ້ແມ່ນໄດ້ຮັບການດົນໃຈຈາກໂຄງສ້າງແລະຫນ້າທີ່ຂອງສະຫມອງຂອງມະນຸດແລະຖືກນໍາໃຊ້ສໍາລັບວຽກງານເຊັ່ນ: ການຈັດປະເພດຮູບພາບແລະການຮັບຮູ້ສຽງເວົ້າ.
- ສະຫນັບສະຫນູນ Vector Machines (SVMs): ປະເພດຂອງແບບຈໍາລອງນີ້ແມ່ນໃຊ້ສໍາລັບບັນຫາການຈັດປະເພດແລະເຮັດວຽກໂດຍການຊອກຫາເຂດແດນທີ່ດີທີ່ສຸດທີ່ແຍກຊັ້ນໃນຂໍ້ມູນ.
- ແບບຈໍາລອງກຸ່ມ: ແບບຈໍາລອງປະເພດນີ້ແມ່ນໃຊ້ສໍາລັບການຈັດກຸ່ມຈຸດຂໍ້ມູນທີ່ຄ້າຍຄືກັນຮ່ວມກັນ ແລະຖືກນໍາໃຊ້ທົ່ວໄປສໍາລັບການແບ່ງສ່ວນຕະຫຼາດ ແລະການສ້າງໂປຣໄຟລ໌ລູກຄ້າ.
- ກົດລະບຽບການສະມາຄົມ: ຮູບແບບປະເພດນີ້ແມ່ນນໍາໃຊ້ສໍາລັບການວິເຄາະກະຕ່າຕະຫຼາດແລະຖືກນໍາໃຊ້ເພື່ອຊອກຫາຄວາມສໍາພັນລະຫວ່າງລາຍການໃນຊຸດຂໍ້ມູນ.

23. CRISP-DM ຫຍໍ້ມາຈາກ

CRISP-DM: Cross-Industry Standard Process for Data Mining

24. ຂໍ້ເສຍຂອງ Clustering

- ການກໍານົດຈໍານວນກຸ່ມ: ມັນສາມາດເປັນການຍາກທີ່ຈະກໍານົດຈໍານວນກຸ່ມທີ່ເຫມາະສົມທີ່ສຸດສໍາລັບຊຸດຂໍ້ມູນທີ່ກໍານົດໄວ້
- ຄວາມລໍາອຽງເບື້ອງຕົ້ນ: ການແກ້ໄຂສຸດທ້າຍທີ່ໄດ້ຮັບຈາກວິທີການຈັດກຸ່ມສາມາດໄດ້ຮັບອິດທິພົນຢ່າງຫຼວງຫຼາຍໂດຍຈຸດເລີ່ມຕົ້ນເບື້ອງຕົ້ນທີ່ເລືອກ.
- ຄວາມອ່ອນໄຫວຕໍ່ກັບ Outliers: ສຸດການຄິດໄລ່ຂອງກຸ່ມສາມາດມີຄວາມອ່ອນໄຫວຕໍ່ກັບ outliers, ແລະພວກມັນອາດຈະສົ່ງຜົນກະທົບຕໍ່ການແກ້ໄຂສຸດທ້າຍ.
- ຮູບຮ່າງຂອງກຸ່ມທີ່ແຕກຕ່າງກັນ: ສຸດການຄິດໄລ່ຂອງກຸ່ມຖືກອອກແບບເພື່ອຈັດການກຸ່ມທີ່ເປັນຮູບຊົງກົມ ຫຼືຮູບກົມ, ແຕ່ພວກມັນອາດຈະບໍ່ເຮັດໄດ້ດີກັບກຸ່ມທີ່ມີຮູບຮ່າງທີ່ແຕກຕ່າງກັນ.
- ຄວາມຫຍຸ້ງຍາກໃນການຕີຄວາມຫມາຍຜົນໄດ້ຮັບ: ຜົນໄດ້ຮັບຂອງກຸ່ມບາງຄັ້ງອາດຈະຍາກທີ່ຈະຕີຄວາມຫມາຍແລະເຂົ້າໃຈ.
- ຄວາມຖືກຕ້ອງຂອງກຸ່ມ: ມັນສາມາດເປັນການທ້າທາຍໃນການກໍານົດຄວາມຖືກຕ້ອງຂອງກຸ່ມແລະການປະເມີນຄຸນນະພາບຂອງກຸ່ມທີ່ໄດ້ຮັບ.
- ຜົນໄດ້ຮັບທີ່ບໍ່ແມ່ນການກໍານົດ: ບາງ algorithms ຂອງກຸ່ມອາດຈະຜະລິດຜົນໄດ້ຮັບທີ່ແຕກຕ່າງກັນໃນແຕ່ລະຄັ້ງທີ່ເຂົາເຈົ້າດໍາເນີນການ, ເຊິ່ງເຮັດໃຫ້ມັນຍາກທີ່ຈະຜະລິດຄືນຜົນໄດ້ຮັບແລະ ໄດ້ຮັບຜົນໄດ້ຮັບທີ່ສອດຄ່ອງ.
- ຄວາມຫຍຸ້ງຍາກໃນການຈັດການຊຸດຂໍ້ມູນຂະຫນາດໃຫຍ່: algorithms ການຈັດກຸ່ມສາມາດມີລາຄາແພງໃນການຄິດໄລ່, ໂດຍສະເພາະແມ່ນສໍາລັບຊຸດຂໍ້ມູນຂະຫນາດໃຫຍ່.

25. Classification ມີຂັ້ນຕອນດັ່ງນີ້

- ສໍາເລັດການປະເມີນຄວາມສ່ຽງຂອງຂໍ້ມູນທີ່ລະອຽດອ່ອນ.

- ພັດທະນານະໂຍບາຍການຈັດປະເພດທີ່ເປັນທາງການ.

- ຈັດປະເພດປະເພດຂອງຂໍ້ມູນ.

- ຄົ້ນພົບສະຖານທີ່ຂອງຂໍ້ມູນຂອງທ່ານ.

- ກໍານົດແລະຈັດປະເພດຂໍ້ມູນ.

- ເປີດໃຊ້ການຄວບຄຸມ.

- ຕິດຕາມກວດກາແລະຮັກສາ.

26. ການຂຸດຄົ້ນຂໍ້ມູນແມ່ນຫຍັງ?

ມັນເປັນຂະບວນການສ້າງພາບພິດຫຼື blueprint ທີ່ກໍານົດລະບົບການເກັບກໍາຂໍ້ມູນແລະການຄຸ້ມຄອງຂອງທຸກອົງການຈັດຕັ້ງ. ແຜນຜັງຫຼືຮູບແບບຂໍ້ມູນນີ້ຊ່ວຍໃຫ້ຜູ້ມີສ່ວນກ່ຽວຂ້ອງເຊັ່ນນັກວິເຄາະຂໍ້ມູນ. ນັກວິທະຍາສາດແລະວິສະວະກອນເພື່ອສ້າງທັດສະນະລວມຂອງຂໍ້ມູນຂອງບໍລິສັດ. ຮູບແບບແນະນໍາສິ່ງທີ່ທຸລະກິດຄວນເກັບກໍາຂໍ້ມູນ.

27. ຄວາມໝາຍຂອງການປະເມີນຄ່າແບບຈຸດ

ຂະບວນການຊອກຫາຄ່າປະມານຂອງບາງພາລາມິເຕີ—ເຊັ່ນ: ຄ່າສະເລ່ຍ (ສະເລ່ຍ)—ຂອງປະຊາກອນຈາກຕົວຢ່າງຂອງປະຊາກອນແບບສຸ່ມ.

28. ຂັ້ນຕອນການທົດສອບນີ້ເຮົາຈະໄດ້ຜົນການວິເຄາະຂໍ້ມູນດ້ວຍເທັກນິກທາງການຂຸດຄົ້ນຂໍ້ມູນ ແລ້ວ ແຕ່ກ່ອນທີ່ຈະເອົາຜົນທີ່ໄດ້ຮັບໄປໃຊ້ງານຕໍ່ໄປນັ້ນ, ເຮົາກໍ່ຈະຕ້ອງມີການວັດປະສິດທິພາບ ຂອງຜົນ

ກ່ອນທີ່ຈະນໍາໃຊ້ຜົນໄດ້ຮັບຂອງການວິເຄາະການຂຸດຄົ້ນຂໍ້ມູນໃນການຕັດສິນໃຈ, ມັນເປັນສິ່ງສໍາຄັນທີ່ຈະວັດແທກປະສິດທິຜົນຂອງພວກເຂົາ. ນີ້ກ່ຽວຂ້ອງກັບການນໍາໃຊ້ວິທີການປະເມີນຜົນຕ່າງໆເພື່ອກໍານົດຄວາມຖືກຕ້ອງແລະຄວາມຫນ້າເຊື່ອຖືຂອງຜົນໄດ້ຮັບ. ວິທີການປະເມີນທົ່ວໄປປະກອບມີການປະເມີນຄວາມຖືກຕ້ອງ, ການປະເມີນຄວາມແມ່ນຍໍາແລະການເອີ້ນຄືນ, ການປະເມີນເສັ້ນໂຄ້ງ ROC, ແລະການປະເມີນຕາຕະລາງຍີກ. ວິທີການປະເມີນທີ່ເຫມາະສົມຈະຂຶ້ນກັບເປົ້າຫມາຍແລະລັກສະນະຂອງໂຄງການຂຸດຄົ້ນຂໍ້ມູນ.